TREBALL DE FI DE MÀSTER

# Màster Interuniversitari en Estadística i Investigació Operativa UPC-UB

**Títol: La regressió quantílica per a les mesures de risc.**

**Autor: Albert Pitarque Méndez**

**Directora: Montserrat Guillen i Estany**

**Departament: Departament d'econometria**

**Universitat: Universitat de Barcelona**

**Convocatòria: Juny, 2019**



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

UNIVERSITAT DE BARCELONA

# Agraïments

La realització d'aquest treball no hagués estat possible sense l'ajut de diverses persones i m'agradaria fer una menció especial a totes elles.

En primer lloc vull agrair-li a la professora Montserrat Guillen tota l'ajuda que m'ha proporcionat durant la realització del treball buscant informació que em podria ser útil, corregint les errades que he tingut i explicant-me en tot moment qualsevol cosa que fos necessària per obtenir els resultats. També li agraeixo la possibilitat que em va donar accedint a ser la meva tutora en el treball de final de màster i introduint-me en els diversos usos de la regressió quantílica. Sens dubte aquest treball hagués estat impossible sense la seva ajuda.

També vull agrair als meus pares, Joan i Anna i al meu germà Joan, tot el suport que m'han donat al llarg de tota la meva carrera acadèmica animant-me a seguir treballant i estudiant per aconseguir complir els meus objectius. Sempre han estat al meu costat en els bons moments i, el que és més important, en els mals moments. No podria ser on soc sense ells. Una altra persona que ha estat important en l'hora d'escollir la direcció que han pres els meus estudis ha estat la professora Isabel Serra de l' Universitat Autònoma de Barcelona que em va introduir en el món del risc i va ser qui, indirectament, em va presentar a la Montserrat.

Per últim vull agrair a la Universitat Politècnica de Catalunya i a la Universitat de Barcelona per haver-me admès en el màster i proporcionar-me la possibilitat de continuar estudiant una cosa que m'apassiona com és l'estadística.

# Índex

# 1 Conceptes prèvis

## 1.1 Definició de quantil

Es diu que un estudiant puntua al quantil $\tau$ d'un examen si ho fa millor que el $\tau\%$ de la resta d'estudiants però pitjor que el $(1-\tau)$. D'una manera semblant el quartils divideixen la població en quatre segments amb igual proporció de la població a cada segment. Els decils ho fan en 10 parts i finalment els quantils o també anomenats percentils es refereixen al cas més general. Dit d'una altre manera, el valor del quantil $\tau$ és aquell valor $c_\tau$ d'una variable aleatoria contínua X que fa que la probabilitat $P(X \leq c_\tau) \leq \tau$ on $c_\tau$ és el valor del quantil $\tau$.

## 1.2 Definició de TVaR

El TVaR o *Tail Value at Risk* és una mesura del risc que quantifica quina és la pèrdua de diners esperada si es sap que és més gran que un valor de pèrdua determinat. El TVaR és conegut per altres noms tals com CTE (Conditional Tail Expectation) o TCE (Tail Conditional Expectation). El TVaR està definit utilitzant un nivell de confiança $\tau$ que es troba entre 0 i 1 que fa referència al quantil al qual s'està calculant aquesta mesura del risc. Escrit de forma matemàtica,

$$TVaR_\tau = E(X|X > c_\tau) \tag{1}$$

on X és la variable que mesura la pèrdua de diners i $c_\tau$ és el valor del quantil $\tau$ en la variable que s'estigui estudiant. Aquesta mesura del risc és freqüentment utilitzada en el camp de les finances per tal de preveure la quantitat de diners que es perdria en cas que passés alguna cosa perjudicial a l'empresa i així poder guardar una part de diners per tal de cobrir-se en cas de necessitat. En aquest àmbit, normalment, es treballa amb les pèrdues i s'agafa una variable aleatoria continua que indica els guanys i pèrdues de l'empresa sent els guanys els valors negatius de la distribució i les pèrdues els valors positius. Sabent això, per calcular el TVaR, s'ha de calcular el valor de la integral:

$$TVaR_\tau = \frac{1}{1-\tau}\int_\tau^\infty xf(x)dx \tag{2}$$

on f(x) és la funció de densitat de la variable X que defineix pèrdues comentada amb anterioritat.

## 1.3 Els orígens de la regressió quantílica

El concepte de regressió és molt antic i actualment la definició més utilitzada indica que és que és un model matemàtic que pretén trobar una relació entre una variable d'interés anomenada variable resposta i un conjunt de variables observables anomenades variables explicatives. Com aquesta relació no és exacta, hi ha una part anomenada residu que és la diferència entre el que s'observa a la variable resposta i l'obtingut mitjançant el model matemàtic que combina les variables explicatives.

El concepte de regressió ha anat evolucionant amb el pas del temps però ho ha fet molt en les últimes decades fins a arribar al concepte actual. El primer en parlar de regressió va ser ser Roger Joseph Boscovich en el segle XVIII. Aquest físic i matemàtic, parlava d'un tipus de regressió bivariant i per estimar-la, restringia la mitjana dels residus de la regressió a ser zero i per tal d'estimar l'efecte d'una variable sobre l'altra, intentava minimitzar la suma dels residus en valor absolut. Aquest tipus de regressió va ser conegut com LAD (Least Absolute Deviations). Una mica més tard Pierre-Simon Laplace proposava utilitzar sistemes d'equacions per tal d'estimar l'efecte de més d'una variable.

Anys més tard que els dos anteriors, al 1795, Carl Friedrich Gauss va descobrir que s'obtenien bons resultats utilitzant el mateix procediment que Laplace però utilitzant la suma dels residus al quadrat en comptes del valor absolut, però aquest procediment no va ser públic fins que a l'any 1806, Adrien-Marie Legendre realitzava un estudi independent amb el qual feia la publicació de l'anomenat "mètode dels mínims quadrats"on s'utilitzaven sistemes d'equacions per tal d'estimar els efectes de les variables sent una continuació de l'estudi de Laplace. Aquest mètode és el que es desenvoluparia fins al mètode dels mínims quadrats ordinaris que es coneix avuí en dia. Tres anys més tard, Gauss va publicar els seus descobriments fent que es generés controversia per qui havia fet el descobriment primer. És interessant que en realitat el métode dels mínims quadrats va ser desenvolupat per tal d'intentar explicar el moviment de planetes i de cometes.

Tornant a la regressió que proposava Laplace, Miles Edgeworth al 1888 va proposar eliminar la restricció que fixava la mitjana dels residus a zero i proposava que s'estimés tant l'efecte d'una variable sobre l'altra com un efecte constant minimitzant la suma de valors absoluts dels residus. Edgeworth va proporcionar un algorítme per a realitzar l'estimació que s'anticipava als mètodes d'optimització actuals anomenats simplex. La proposta d'Edgeworth va anar perdent força fins que als anys 50 va ser recuperada i va ser reconeguda com un problema de programació lineal.

Tot i que els dos mètodes estimaven els efectes de les variables de forma eficient, es va observar que l'aproximació per la suma de residus en valor absolut tenia l'avantatge que era més resistent quan hi havia

presència de valors especialment grans en les variables. Però aquest mètode tenia el problema que degut a què no hi havia un coneixement extens dels mètodes de computació, no es disposava de la capacitat de realitzar inferència estadística.

La regressió quantílica que pretén modelitzar el quantil d'una variable resposta en funció d'una combinació lineal de regressors, té els orígens en el mètode LAD. Un dels primers estudis que es va relitzar utilitzant la regressió quantílica va ser per fer regressió a la mediana és a dir al quantil 50. Aquest estudi va ser realitzat per KJ.Arrow i M.Hoffenberg l'any 1959 i estava relacionat amb la producció i distribució de productes de diverses indústries. Utilitzaven l'aproximació per la suma de valors absoluts de la desviació.

Fins al 1978, els estudis que utilitzaven la regressió quantílica es centraven en fer regressió per a la mediana. Uns dels primers en estudiar la regressió per a diferents quantils van ser Gilbert Bassett i Roger Koenker al 1978 i les seves propostes han anat evolucionant els darrers anys fins a obtenir la regressió quantílica que es coneix ara i que s'ha començat a fer servir amb força en molts àmbits.

## 1.4 Estudis recents

| Autors (any) | Títol i revista | Conclusions | Dades | Mètodes addicionals |
|---|---|---|---|---|
| Kaza, N. (2010) | Understanding the spectrum of residential energy consumption: a quantile regression approach. *Energy policy*, 38(11) 6574-6585 | Utilitzant la regressió quantílica s'estudia com afecten diverses característiques de cases en l'estalvi energètic i el consum d'energia. | N = 4.382 K = 9 $\tau$ = (0.1, 0.3, 0.5, 0.7, 0.9) | OLS |
| Behr, A. (2010). | Quantile regression for robust bank efficiency score estimation. *European Journal of Operational Research*, 200(2), 568-581. | Estima l'eficiència de diferents tipus de bancs. Compara l'eficiència dels bancs entre ells. Compara els resultats obtinguts de la regressió quantílica amb altres tipus d'anàlisis d'eficiència. | N = 1.942 K = 10 $\tau$ de 0 a 1 | Simulacions de Monte-Carlo. SFA Value-Added functions |
| Liao, W. C., & Wang, X. (2012). | Hedonic house prices and spatial quantile regression. *Journal of Housing Economics*, 21(1), 16-27. | Utilitza la regressió quantílica per ajustar preus de vivendes a una ciutat emergent a Xina. Es troba que hi ha dependència entre el quantil del preu i la localització. | N = 46.356 K = 14 $\tau$ de 0 a 1 | Economia espacial, OLS |
| Tareghian, R., & Rasmussen , P. F. (2013). | Statistical downscaling of precipitation using quantile regression. *Journal of hydrology*, 487, 122-135. | Fan predicció de les plujes diaries a diferents punts d'Estats Units. Comparen els resultats obtinguts amb els mètodes que s'utilitzaven al 2013 per analitzar aquest tipus de variables. | N = 27.375 aprox K = 14 $\tau$ de 0 a 1 | Global Climate Models, OLS. Es proposen més articles relacionats amb la meteorologia |

| Autors (any) | Títol i revista | Conclusions | Dades | Mètodes addicionals |
|---|---|---|---|---|
| Daniel-Spiegel, E., Weiner, E., Yarom, I., Doveh, E., Friedman, P., Cohen, A., & Shalev, E. (2013). | Establishment of fetal biometric charts using quantile regression analysis. *Journal of Ultrasound in Medicine*, 32(1), 23-33. | S'aplica la regressió quantílica per detectar anormalitats en el creixement de fetus durant l'embaraç. Es centren en la circumferència del cap, la circumferència del abdomen i la llargada del fèmur. | N = 11.169<br>K = 5<br>$\tau$ de 0 a 1<br><br>(Només utilitzen una variable a la que han aplicat transformacions) | Hardlock OLS |
| Briollais, L. & Durrieu, G. (2014) | Application of quantile regression to recent genetic and omic studies. *Human genetics, 133(8), 951-966* | Resum d'aplicacions de la regressió quantílica a genètica i ciències òmiques. Fa una extensió de la regressió quantílica. Cas aplicat i software en R. | N = 378<br>K = 6<br>$\tau$ = 0.5 | OLS i còpula-regressió quantílica |
| Valenzuela, C. Valencia, A., White, S., Jordan, J. A., Cano, S., Keating, J., ... & Potter, L. B. (2014). | An analysis of monthly household energy consumption among single-family residences in Texas, 2010. *Energy Policy*, 69, 263-272. | S'utilitzen variables demogràfiques, socio-econòmiques i característiques de les cases per estudiar el consum d'energia i desenvolupar estratègies per estalviar energia. | N = 294.416<br>K = 18<br>$\tau$ = (0.1, 0.5, 0.9) | |
| Marrocu, E., Paci, R., & Zara, A. (2015). | Micro-economic determinants of tourist expenditure: A quantile regression approach. *Tourism Management*, 50, 13-30. | Ajusta diversos models de regressió per a la despesa de diners dels turistes que van anar a Sardenya l'any 2010. | N = 1.445<br>K = 16<br>$\tau$ de 0 a 1 | OLS |
| Niemierko, R. Töppel, J. & Tränkler, T. (2019) | A D-vine copula quantile regression approach for the prediction of residential heating energy consumption based on historical data. *Applied Energy*, 233, 691-708. | Analitzant la distribució sencera del consum energètic en funció de les característiques de l'edifici es comprova que els efectes són diferents als diferents quantils. | N = 25.000 llars alemanyes<br>K = 16<br>$\tau$ de 0 a 1 | Artificial Neural Networks i còpula-regressió quantílica<br><br>Fa 116 referències a articles anteriors |

# 2 Motivació i objectius

La regressió quantílica és una eina que pot dir-se que ha estat eclipsada per la regressió lineal i altres tipus de tècniques d'ajust de models per a explicar una resposta amb diversos tipus de dades. Ha estat en els últims anys que aquest tipus de regressió ha anat evolucionant i cobrant importancia però tot i el gran nombre d'avanços, encara queden moltes aplicacions d'aquesta per desenvolupar. Es per això que en aquest treball s'ha intentat aplicar aquesta técnica d'ajust per a estimar mesures de risc que és una possible aplicació i extensió.

El principal objectiu d'aquest treball és proposar un model que permeti ajustar la mesura del risc coneguda com a Tail Value at Risk (TVaR) o anomenada també Conditional Tail Expectation (CTE) o Tail Conditional Expectation (TCE). Aquesta mesura del risc és una extensió d'una altra mesura del risc anomenada Value at Risk (VaR) que és molt utilitzada sobretot en el camp de l'economia i que correspón a la noció de quantil. El problema principal que té aquesta mesura del risc és que només es té en compte el valor en un quantil en concret i això pot portar a perdre una gran quantitat d'informació sobre els valors que es troben per sobre d'aquest quantil. En canvi el TVaR si que té en compte aquests valors per sobre del quantil i permet una millor graduació del risc o si més no una visió diferent.

D'altra banda, un repte important en l'elaboració d'aquest treball és el software que s'ha utilitzat per dur-lo a terme. Moltes empreses utilitzen softwares de pagament tals com SAS o SPSS perquè hi ha una revisió molt detallada i es garanteix que les funcions que hi ha programades en el software siguin correctes. Aquests programes són molt cars per als estudiants i per a algunes petites empreses. És per això que s'ha decidit elaborar aquest treball en el software estadístic R que és completament gratuït. Actualment l'R té un paquet de funcions per ajustar la regressió quantílica però, lamentablement, aquest paquet no disposa de cap funció per ajustar al TVaR per tant se n'ha hagut de crear una de nova. En realitat un objectiu del treball era precusament trobar un algoritme per a la regressió del TVaR.

La regressió quantílica s'ha aplicat en dos casos amb dades reals per tal d'ajustar diversos quantils i posteriorment per ajustar el TVaR per als mateixos quantils. Aquestes dues bases de dades que s'han seleccionat per a realitzar aquest treball tenen relació amb temes on en l'actualitat s'utilitzen les mesures de risc de forma habitual.

La primera base de dades està relacionada amb el consum d'energía de diverses llars als Estats Units. En aquesta base de dades s'especifiquen diverses característiques sobre algunes cases tals com la superfície, el nombre d'habitants de la casa, el nombre de telèfons mòbils o si tenen piscina o no d'entre d'altres. L'objec-

tiu utilitzant aquesta base de dades és trobar un model que ajusti el TVaR per a la despesa d'energia en dòlars.

La segona base de dades està relacionada amb assegurances de cotxes. En els darrers anys, les asseguradores han començat a realitzar ofertes amb descomptes per als bons conductors. Per tal de saber si els solicitants són o no bons conductors recullen constantment dades mitjançant un aparell que es posa en el cotxe. L'objectiu utilitzant aquesta base de dades és que utilitzant variables com la zona per la qual es sol conduir o a quina hora es pugui ajustar el TVaR per al nombre de kilometres que es condueix per sobre del límit de velocitat.

# 3 Metodologia: Regressió quantílica

La tècnica que s'ha utilitzat principalment en aquest estudi és la regressió quantílica que és un tipus de regressió que permet trobar la relació entre els quantils d'una variable resposta amb un conjunt de variables explicatives. A continuació s'explica com funciona la regressió quantílica i com es pot aplicar en el programa R.

## 3.1 Introducció a la regressió quantílica

Abans de parlar sobre la regressió quantílica cal parlar sobre la regressió per mínims quadrats ordinaris ja que la regressió quantílica és una extensió d'aquesta.

La regressió per mínims quadrats ordinaris és un mètode que serveix per estimar els coeficients de la regressió lineal. La regressió lineal és una aproximació lineal per modelar la relació entre una variable resposta (o variable dependent) i una o més variables explicatives (o variables independents). La regressió lineal, normalment, s'utilitza per modelar la mitjana de la variable resposta però també es pot utilitzar per modelar quantils determinats o la mediana. La relació que es troba utilitzant aquest tipus de regressió té l'equació següent:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i, \tag{3}$$

on Y és el valor de la variable resposta, $X_i sónelsvalorsdelesvariablesexplicativesi\beta$ representa l'efecte que té cada variable explicativa a la variable resposta. $\beta_0$ és un efecte constant i no representa a cap variable sinó a un terme constant. Aquestes $\beta$ es denominen coeficients. Per últim, $\epsilon_i$ representa la diferència que hi ha entre la relació calculada a partir dels coeficients i les variables explicatives i el valor de la variable resposta. Aquesta diferència s'anomena residu i ha de complir la condició $E(\epsilon_i|X = X_i) = 0$.

Per la hipòtesi anterior es pot deduir que $E(Y|X = X_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$ i per tant que l'esperança matemàtica de la resposta de l'i-èssima observació és igual a la combinació lineal dels seus regressors.

El mètode dels mínims quadrats ordinaris és el mètode més utilitzat per calcular el valor dels coeficients. Aquest mètode es basa en calcular els coeficients $\beta$ de tal manera que la suma de quadrats de la diferència entre els valors de la variable resposta i els valors predits a partir de les variables explicatives i els coeficients,

sigui el mínim.

**Definició 1**

Sigui X la matriu de les variables explicatives, Y el vector de la variable resposta i $\beta$ el vector de coeficients, es té que:

$$\mathrm{X} = \begin{pmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,n} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & X_{m,1} & X_{m,2} & \cdots & X_{m,n} \end{pmatrix} \quad \mathrm{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}$$

El mètode dels mínims quadrats ordinaris calcula $\hat{\beta}$ de tal manera que

$$\hat{\beta} = arg_{\beta} min S(\beta) \tag{4}$$

on

$$S(\beta) = ||Y - X\beta||^2 \tag{5}$$

Un cop introde-camuida la regressió lineal i el mètode dels mínims quadrats ordinaris, ja es pot començar a parlar sobre conceptes més relacionats amb la regresió quantílica. La regressió quantílica és un tipus d'anàlisi de regressió que té com a objectiu estimar la mediana o alguns quantils concrets de la variable resposta a diferència de la regressió lineal que ho feia per la mitjana. Aquest tipus de regressió és utilitzat sobretot quan hi ha presència d'outliers a la variable resposta proporcionant una estimació dels coeficients més robusta que la regressió lineal. No hi ha diferència entre els dos mètodes en la qualitat de l'estimació pel fet de que hi hagi presència d'outliers a les variables explicatives

Per tal de comprovar que la regressió quantílica és millor quan hi ha presència d'outliers a la variable resposta que la regressió lineal s'estudia una funció anomenada funció d'influència. La funció d'influència és una funció que descriu l'efecte que té una observació en la estimació d'un estadístic. La funció d'influència

es defineix:

$$IF(y, \hat{\gamma}, F) = \lim_{t \to 0} \frac{\hat{\gamma}(F_t) - \hat{\gamma}(F)}{t} \tag{6}$$

Mitjançant l'estudi de com es comporta la funció d'influència (Veure més detalls a la referència [1]), es comprova com influeixen les diferents observacions d'una Normal(0,1) en la estimació de la mitjana i de la mediana.

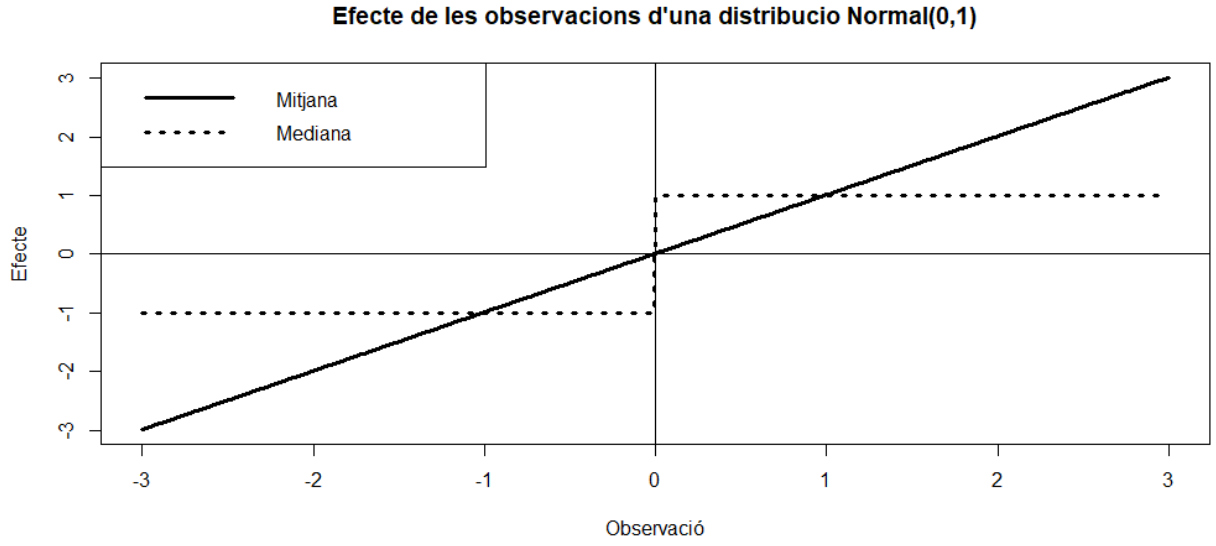


**Figura 1:** Efecte de les observacions d'una Normal(0,1) en l'estimació de la mitjana i la mediana

Com s'observa en el gràfic de la Figura 1 per a l'estimació de la mitjana, la influència que tenen els valors és proporcional al valor de la observació per tant els valors més extrems o valors anomals tindran més influència a l'hora de realitzar l'estimació per a la mitjana. En canvi, en l'estimació de la mediana, les observacions tenen el mateix efecte en valor absolut i aquest efecte no es veu afectat per els valors extrems de la variable.

Abans de pasar a parlar de l'estimació dels coeficients de la regressió quantílica en més profunditat s'ha de definir la funció de pèrdua del quantil. Tenint en compte la definició de quantil realitzada al primer capítol, es defineix la funció pèrdua del quantil amb la següent formulació:

$$\rho_\tau(u) = u(\tau - I(u < 0)), \tag{7}$$

on u és una constant, $\tau$ és el quantil sobre el que s'està mirant la funció pèrdua i I és una funció indicadora. Les funcions indicadores valen 1 quan es compleix la condició especificada i 0 si no es compleix. En aquest

cas la condició és u $< 0$. Aquesta funció pèrdua, com es pot comprovar en el gràfic de la Figura 2, no és diferenciable i això pot portar problemes a l'hora de resoldre problemes d'optimització en les que la funció pèrdua estigui involucrada.

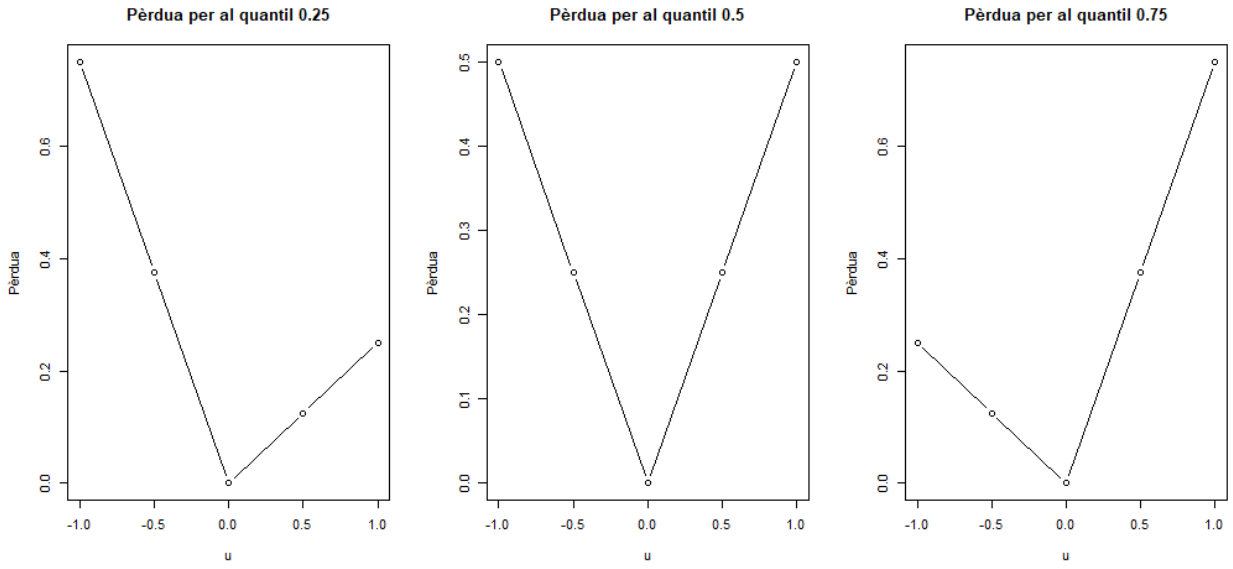


**Figura 2:** Gràfics de les funcions de pèrdua per a diferents quantils

En els gràfics de laFfigura 2 s'han dibuixat tres funcions de pèrdua per als quantils 0.25, 0.50 i 0.75 utilitzant com a valor u els valors -1, -0.5, 0, 0.5 i 1. Com es pot observar en els tres casos la funció té forma de punxa sent u = 0 el valor mínim de la funció. Aquest fet que la funció tingui la forma de punxa és el que provoca que la funció de pèrdua no sigui diferenciable.

Aquesta funció de pèrdua també és útil per calcular quins són els quantils d'una mostra de observacions. Així que donades les observacions $X_1, X_2, ..., X_n$ d'una variable aleatòria X, els quantils de X poden ser definits com:

$$\hat{c}_\tau = arg\min_c \int \rho_\tau(x-c)dF_n(x) = arg\min_c \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(X_i - c) \tag{8}$$

on es segueix la notació utilitzada en el primer capítol i on $F_n(x)$ és la distribució empírica de la variable aleatòria X. Una demostració extensa d'aquesta definició es pot trobar a la secció 1.1.2 del llibre [2]. Aquesta definició serà útil a l'hora d'estimar els efectes de la regressió quantílica.

## 3.2 Regressió quantílica paramètrica

Tot i que, com en la regressió lineal es pot distingir la regressió entre paramètrica i no paramètrica, en aquest treball només s'ha utilitzat la regressió paramètrica. La regressió s'anomena paramètrica quan s'assumeix que amb un conjunt finit de variables es pot definir la distribució de la variable resposta.

### 3.2.1 L'estimador

Es defineix un model de regressió lineal bàsic per tal d'explicar com s'estimen els efectes de la regressió quantílica. Si $c_\tau$ és el quantil de la resposta $Y_i$ de la i-èssima observació, es diu que

$$c_{\tau i} = \beta_\tau X_i + \epsilon_i : \tag{9}$$

En aquest model els residus $\epsilon_i$ han de complir la condició $P(\epsilon_i \leq 0 | X = X_i) = \tau$ que és equivalent a la condició $E(\epsilon_i | X = X_i) = 0$ que es troba en el mètode dels mínims quadrats ordinaris del model lineal. En l'article [3] es proposa que tenint en compte la fòrmula per calcular el quantil de la mostra, sería raonable resoldre el següent problema d'optimització per tal de calcular l'efecte $\beta_\tau$ que mesura l'impacte de $X_i$ sobre el quantil de la resposta.

$$\widehat{\beta_\tau} = \text{argmin}_b \left[ \sum_{Y_i \geq X_i b} \tau |Y_i - X_i b| + \sum_{Y_i < X_i b} (1-\tau) |Y_i - X_i b| \right] \tag{10}$$

Tenint en compte la definició de funció de pèrdua del quantil proporcionada anteriorment, l'equació es pot simplificar en la següent que es tenen en compte K regressors:

$$\widehat{\beta_\tau} = arg_\beta min \sum_{i=1}^{n} \rho_\tau (Y_i - X'_{ji} \beta_\tau \tag{11}$$

que és una extensió del problema d'optimització que s'ha vist per calcular els coeficients $\beta$ de la regressió lineal per mínims quadrats ordinaris. En l'article [4] es demostra que l'estimador és consistent a l'hora d'estimar $\beta_\tau$ és a dir, l'error de l'estimador tendeix a 0 quan la grandaria de la mostra es va fent gran.

A diferència del què es sap que els estimadors del coeficients de la regressió per mínims quadrats ordinaris segueixen una distribució t-Student, per als coeficients de la regressió quantílica no es coneix la distribució exacta dels estimadors. Tot i així, sota certes condicions, s'ha estudiat que $\sqrt{n}(\widehat{\beta_\tau} - \beta_\tau)$ tendeix a seguir una distribució Normal.

**Teorema 1** *Tenint un model lineal com el de la equació (9). Sota les següents condicions:*

- *La funció de distribució condicional $F_i$ de $Y_i$ condicionada a $X_i$ és absolutament continua amb funció de densitat $f_i$ també continua i que per cada quantil $c_\tau$ pren valors entre 0 i $\infty$.*
- *Existeixen les matrius $D_0$ i $D_1(\tau)$ definides positives tals que:*

  1. $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i X_i' = D_0$.
  2. $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f_i(c_i(\tau) X_i X_i') = D_1(\tau)$.
  3. $max_{i=1,\ldots,n} ||X_i|| / \sqrt{n} \to 0$.

  *es verifica que*

$$\sqrt{n}(\widehat{\beta_\tau} - \beta_\tau) \xrightarrow{d} N(0, \tau(1-\tau) D1(\tau)^{-1} D_0 D_1(\tau)^{-1}),$$

La demostració d'aquest teorema es troba detallada a l'article [1].

### 3.2.2 Càlcul dels errors estàndard

S'han investigat molts mètodes per calcular els errors estàndard en la regressió quantílica però un dels més utilitzats i el que s'ha utilitzat en aquest treball pel fet de ser molt fàcil d'aplicar és el metode bootstrap.

El bootstrap és un mètode de remostreig que s'utilitza per estimar certes característiques dels estimadors, com per exemple la variància, calculant-les a partir d'una distribució aproximada. Una distribució utilitzada normalment és la distribució empírica de les dades Aquest mètode també serveix per a construir tests d'hipòtesi i s'utilitza com alternativa a l'estadística inferencial. Amb l'aparició dels ordinadors ha augmentat molt la qualitat i rapidesa d'aquest mètode ja que aquests permeten remostrejar moltes vegades a gran velocitat. El procediment per dur a terme aquest mètode és el següent:

1. El primer pas consisteix en trobar de forma aleatòria una de les submostres. Els programes amb els què es treballa per fer aquest tipus de procediments ja tenen incorporades comandes amb les quals es poden obtenir submostres aleatòries de la mida desitjada a partir d'una mostra més gran.
2. Un cop trobada la submostra, es calcula l'estadístic del qual es volen obtenir les propietats i es torna a trobar una nova submostra aleatòria per tal de tornar a calcular l'estadístic. Aquest procediment es repeteix tantes vegades com possibles submostres hi hagi o, en el seu defecte, un nombre suficientment gran de vegades.

3. Per últim, un cop s'han calculat els estimadors de totes les submostres aleatòries, es calcula la propietat desitjada i aquesta es considera un bon estimador.

Es planteja un exemple molt simple per tal d'entendre millor el procediment del mètode bootstrap. Es suposa que es té una mostra de 20 observacions sobre les alçades de diferents persones i es vol saber quina seria la desviació típica de la estimació de la mitjana de les alçades. Les alçades en centímetres són les següents:

| Alçades de 20 persones | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 162 | 167 | 174 | 165 | 162 | 174 | 163 | 177 | 168 | 172 |
| 175 | 182 | 177 | 148 | 169 | 151 | 170 | 168 | 161 | 165 |

**Taula 1:** Dades d'exemple per al mètode bootstrap

Per calcular aquesta desviació típica utilitzant el mètode bootstrap, el que es farà és calcular la mitjana d'una submostra aleatòria utilitzant només 19 de les observacions per tant hi ha 20 possibles submostres aleatòries. Com que 20 és un nombre suficientment baix, es calcularia la mitjana de les 20 submostres i es calcularia la desviació típica de les mitjanes.

En aquest treball s'ha utilitzat el mètode bootstrap per calcular les desviacions típiques de les estimacions dels coeficients de la regressió quantílica ja que no s'ha utilitzat la funció definida a l'R per ajustar la regressió quantílica. S'han agafat 1000 submostres aleatories de les dades que han estat de 3000 observacions en la base de dades del consum energètic i de 6000 en la base de dades de les assegurances. D'aquestes submostres s'han calculat els valors de les $\beta$ de cada variable i s'ha calculat la desviació típica utilitzant totes elles.

## 3.3 Contrasts d'hipòtesis

Tant com a la funció que s'ha creat per tal d'ajustar la regressió quantílica com en la que ja ve programada en el programa que s'ha fet servir per a realitzar aquest treball, apareixen el valor d'un estadístic t i un p-valor. En aquesta part del treball es vol fer una breu explicació sobre què és un contrast d'hipòtesi i quin és el que s'ha fet servir en aquest treball.

Un contrast d'hipòtesi és un procediment estadístic que permet comprovar si un paràmetre de la població que s'està estudiant compleix alguna característica en concret. Un contrast d'hipòtesi parteix de dues hipòtesis. La hipòtesi nul·la denotada $H_0$ sol ser la hipòtesi que suposa que el paràmetre que es vol estudiar compleix amb la característica. La hipòtesi alternativa, denominada $H_1$ és el contrari de la hipòtesi nul·la.

Per dur a terme el contrast s'ha de calcular un estadístic que sovint s'anomena estadístic t.

El contrast d'hipòtesi més comú i el que s'ha utilitzat en aquest treball pren com a hipòtesi nul·la $H_0 : \hat{\beta} = 0$ i hipòtesi alternativa $H_1 : \hat{\beta} \neq 0$ tot i que en comptes de que $\hat{\beta}$ sigui igual a 0 també pot ser igual a qualsevol altre valor. Per saber si es pot acceptar o rebutjar la hipòtesi nul·la cal calcular un estadístic que depén del contrast que es vulgui realitzar. En aquest cas l'estadístic t s'ha calculat amb la formula:

$$t = \frac{\hat{\beta} - \beta}{\sigma} \tag{12}$$

on $\hat{\beta}$ és el paràmetre que es vol estudiar, $\beta$ és el valor sobre el qual es vol comparar l'estimació realitzada, que en el cas d'aquest contrast d'hipòtesi val 0 i $\sigma$ és la desviació típica del paràmetre. Un cop s'ha calculat el valor de l'estadístic t, es suposa que en una mostra gran aquest segueix una distribució Normal(0,1) i es mira quina és la probabilitat de que hi hagi un valor superior en valor absolut al de l'estadístic t. Aquesta probabilitat s'anomena p-valor i és el que permet determinar quina de les hipòtesis plantejades s'accepta. Normalment es considera que es rebutja la hipòtesi nul·la si el p-valor és inferior a un nivell de significació determinat entre 0 i 1 i no es rebutja si és superior. Usualment, aquest nivell de significació és de 0.05 però en alguns àmbits com per exemple en el camp de l'economia, es pot requerir d'un nivell de significació del 0.01. Cal destacar que hi ha molt contrasts d'hipòtesi possibles i que la forma de calcular l'estadístic i el p-valor pot variar però en aquest treball s'ha explicat el cas més comú.

# 4 Metodologia proposada per estimar el TVaR

## 4.1 Introducció

Comunament les mesures del risc s'utilitzen per a mesurar el risc en el camp de les finances. La mesura del risc utilitzada més freqüent és el Valor at Risk al nivell $\tau$ ($VaR_\tau$) que no és més que el valor del quantil $\tau$ en la funció de distribució d'una variable aleatòria. Un dels principals problemes d'aquesta mesura del risc és que al prendre el valor del quantil com a mesura del risc, no es tenen en compte els valors de les pèrdues superiors al $VaR_\tau$ i ha estat molt criticada per això.

Com ja s'ha dit a la primera part, diverses vegades, una altra mesura del risc és el Tail Value at Risk ($TVaR_\tau$) o també anomenat Expected Shortfall ($ES_\tau$) o Conditional Value at Risk ($CVaR_\tau$) que ja s'ha definit anteriorment. Aquesta mesura del risc, al ser l'esperança dels valors que superen al $VaR_\tau$ si que té en compte les pèrdues superiors i és una mesura del risc coherent. Tot i així, hi ha certa discussió sobre si el TVaR és una mesura del risc robusta. Per una banda, hi ha autors que defensen que el $VaR_\tau$ es preferible davant el $TVaR_\tau$ ja que consideren que és més robust. Per altra banda, alguns autors consideren que no és necessari tenir en compte la robustesa per a mesures del risc i defensen que el $TVaR_\tau$ és una bona mesura.

Un dels articles que ha servit com a base per a realitzar aquest treball ha estat l'article d'Annals of Statistics [14] publicat l'any 2016 i en ell s'introdueix la propietat matèmatica anomenada elicitabilitat. Aquesta propietat la compleixen algunes mesures del risc i consisteix en què una mesura del risc és elicitable si existeix alguna funció de puntuació que permeti comparar models de risc per veure quin és millor. Aquesta elicitabilitat permet ajustar diversos tipus de regressions i una d'aquestes és la regressió quantílica, El $VaR_\tau$ compleix aquesta propietat d'elicitabilitat mentre que el $TVaR_\tau$ no ho fa. Tot i així, això no vol dir que no es pugui aplicar la regressió quantílica per al TVaR ja que en el mateix article es comenta que hi ha un tipus de mesura del risc que per sí sola no és elicitable però sí que pot ser part d'un conjunt de mesures que si que ho siguin. En aquest article es proposa una manera d'ajustar el VaR i el TVaR a la vegada mitjançant la regressió quantílica però aquest ajust no s'ha seguit en aquest treball.

En una publicació anterior [15], Acerbi i Székely, proposen una funció de puntuació per a la parella de mesures de risk ($VaR_\tau$, $TVaR_\tau$) de tal manera que juntes són elicitables. Tant la funció com la demostració d'aquesta propietat es troben a les publicacions [14] i [15]. Com es pot comprovar aquestes dues publicacions són del 2015 i del 2016 és a dir, són articles molt recents i per tant s'ha avançat poc en aquest camp. En l'article del 2016 només s'esmenta que es pot realitzar regressió quantílica per a les mesures de risc però no s'implementa en cap cas amb dades reals.

Aquest treball, partint de la base dels articles de Koenker sobre la regressió quantílica i havent comprovat mitjançant els articles de Fissler i Ziegel [14] i Acerbi i Székely [15] que es pot aplicar la regressió quantílica sobre mesures de risc, preten aportar un avenç en l'ajust de les mesures de risc utilitzant un model matemàtic obtingut a partir de la regressió quantílica. Això representaria un pas endavant en aquest camp ja que no s'ha realitzat una gran quantitat d'estudis prèvis relacionant aquests dos conceptes.

De la mateixa manera que per a la regressió quantílica aplicada a un quantil, un aspecte important per ajustar el model per al TVaR és la funció de pèrdua del quantil introduida en la secció anterior i que seguirà apareixent en la funció objectiu que caldrà optimitzar. No obstant, tot i tenir la mateixa definició, en el cas de l'ajust per al TVaR, aquesta funció de pèrdua haurà de ser modificada.

## 4.2 Regressió paràmetrica basada en el TVaR

La regressió quantílica paramètrica aplicada al $TVaR : \tau$ és una extensió de la regressió quantílica per a un quantil $\tau$ i per això la manera de calcular quins són els efectes de les variables és bastant similar a l'explicada a la secció 3.2.1 on es parla de l'estimador de la regressió quantílica.

Cal recordar que per tal de calcular els valors dels coeficients de la regressió quantílica per un quantil, era necessari minimitzar la funció objectiu

$$\widehat{\beta_\tau} = arg_\beta min \sum_{i=1}^{n} \rho_\tau(Y_i - X'_{ji}\beta_\tau \tag{13}$$

on la funció de pèrdua era el que feia variar els coeficients de la regressió lineal respecte els coeficients de la regressió quantílica i fent també que hi haguessin canvis en els coeficients per als diferents valors $\tau$. En aquest cas, la regressió quantílica no s'està ajustant per a un quantil en concret sinó que s'està ajustant la regressió per al quantil que correspon a la mitjana dels valors que superen el valor del quantil $\tau$. Aquest quantil del TVaR no és el mateix per a totes les dades i cal fer el càlcul cada vegada.

Partint de la definició de TVaR que s'ha plantejat anteriorment, es pot interpretar el TVaR com una esperança matemàtica i per calcular l'esperança d'una variable aleatoria continua amb funció de densitat f(x), s'ha de fer l'integral multiplicant per x, és a dir:

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx \tag{14}$$

Quan es desitja calcular l'esperança de tota la variable aleatoria, els límits de la integral van desde $-\infty$

fins a $\infty$ però en aquest cas l'interés es troba en l'esperança dels valors que es troben per sobre del valor $\tau$. En la funció objectiu que s'utilitza per a estimar els coeficients per al quantil, s'integra sobre la variable $\tau$ que com que només pot prendre valors entre 0 i 1 farà que els límits de l'integral siguin desde $\tau$ fins a 1. Si es resol aquesta integral s'obté la següent fórmula.

$$\begin{aligned} \int_{\tau}^{1} \tau\rho_{\tau}(Y - X\beta)d\tau = (Y - X\beta)\int_{\tau}^{1} \tau(\tau - I(Y - X\beta) \leq 0)d\tau \\ = (Y - X\beta)\left(\frac{1}{3} - \frac{I((Y - X\beta) \leq 0)}{2} - \frac{\tau^3}{3} + \frac{\tau^2 I((Y - X\beta) \leq 0)}{2}\right) \end{aligned} \tag{15}$$

obtenint que la funció objectiu que cal minimitzar per obtenir els coeficients del model ajustat per al TVaR és:

$$\beta_{\tau}(\widehat{TVaR}) = arg_{\beta}min \sum_{i=1}^{n} \left[(Y_i - X_i'\beta)\left(\frac{1}{3} - \frac{I((Y_i - X_i'\beta) \leq 0)}{2} - \frac{\tau^3}{3} + \frac{\tau^2 I((Y_i - X_i'\beta) \leq 0)}{2}\right)\right] \tag{16}$$

Aquesta funció també serà aplicada per quan es calculin les desviacions típiques de les estimacions dels efectes de les variables. En aquest cas, de la mateixa manera que s'ha fet per a la regressió quantílica per ajustar un quantil, s'ha utilitzat el mètode bootstrap per al qual s'han agafat 3000 observacions per al remostreig de la primera base de dades i 6000 per al remostreig de la segona.

# 5 Bases de dades

En aquest apartat es realitzarà una explicació de les bases de dades utilitzades en aquest estudi. S'han utilitzat dues bases, una relacionada amb el consum d'energia als Estats Units l'any 2015 i l'altra relacionada amb assegurances de cotxes. Les bases de dades estan inicialment en format SAS i són analitzades amb R.

## 5.1 Consum d'energia

Aquesta base de dades ha estat obtinguda de la Residential Energy Consumption Survey (RECS) que és una enquesta americana que recull informació relacionada amb el consum d'energia i les característiques de la llar de primeres residències de la gent a la què s'ha fet l'enquesta. La RECS és una enquesta que es duu a terme cada 6 anys i en el cas d'aquest treball s'han seleccionat les dades corresponents a l'enquesta realitzada a l'any 2015.

La base de dades disposa de 5686 observacions i 739 variables de les quals s'ha fet una selecció de les que s'ha cregut que eren les més importants per tal d'aplicar la regressió quantilica de forma correcta. No s'ha realitzat un estudi en profunditat de les 739 variable per determinar quines serien les variables realment importants però s'ha seguit el criteri de posar al model aquelles variables que els investigadors consideren rellevant. Per això ha estat útil la revisió dels treballs efectuada al final del primer capítol. Les variables seleccionades i la seva descripció es poden trobar a la Taula 2 d'aquest document.

| Tipus | Variable | Descripció |
|---|---|---|
| Dependent | DOLLAREL | Cost en dolars de l'energia gastada el 2015 |
| Independent | TOTSQFT_EN | Superfície de la casa en peus quadrats |
| | SWIMPOOL | Variable binaria que val 1 si la casa té piscina i 0 si no. |
| | TOTCSQFT | Superfície de la casa que no està finalitzada o no és habitable en peus quadrats. Fa referència a parkings, àtics, porxos... |
| | TOTHSQFT | Superfície de la casa que té calefacció en peus quadrats. |
| | BEDROOMS | Nombre de dormitoris. |
| | SOLAR | Variable binaria que val 1 si la casa genera energia solar i 0 si no. |
| | TVCOLOR | Nombre de televisors a la casa |
| | NUMFLOORFAN | Nombre de ventiladors a la casa |
| | NUMSMPHONE | Nombre de smart phones a la casa |
| | NHSLDMEM | Nombre de persones vivint a la casa |

**Taula 2:** Descripció de les variables de la base de dades del consum d'energia.

Es realitza una anàlisi estadística descriptiva sobre les diverses variables utilitzades i més exhaustivament sobre la variable dependent per tal d'avaluar quina seria la tècnica d'ajust a aplicar més adhient. La Taula 3 mostra els estadístics principals de la variable resposta.

| Variable | Mitjana | Desviació Típica | Mediana | Mínim | Màxim | Asimetria | Curtosi |
|---|---|---|---|---|---|---|---|
| DOLLAREL | 1538.36 | 823.37 | 1391.27 | 18.72 | 8121.56 | 1.53 | 7.90 |
| TOTSQFT_EN | 2384 | 1264 | 2102 | 228 | 8501 | 1.21 | 4.79 |
| TOTCSQFT | 1670.47 | 1302.38 | 1522 | 0 | 8066 | 0.99 | 4.41 |
| TOTHSQFT | 2075.71 | 1172.28 | 1847 | 0 | 8066 | 1.21 | 5.29 |
| BEDROOMS | 3.15 | 0.92 | 3 | 0 | 10 | 0.43 | 4.31 |
| TVCOLOR | 2.53 | 1.31 | 2 | 0 | 9 | 0.87 | 4.17 |
| NUMFLOORFAN | 0.83 | 1.16 | 0 | 0 | 14 | 1.92 | 10.25 |
| NUMSMPHONE | 1.68 | 1.34 | 2 | 0 | 8 | 0.68 | 3.45 |
| NHSLDMEM | 2.70 | 1.45 | 2 | 1 | 12 | 1.15 | 4.69 |

**Taula 3:** Estadístics descriptius de la base de dades del consum d'energia

A part dels estadístics descriptius que s'observen a la Taula 3, un 10% de les llars disposen de piscina i només un 1.8% produeixen energia solar. Tornant als valor de la taula, pel que fa a la variable dependent, es pot observar que hi ha més presència de valors elevats que de valors baixos ja que la mediana és bastant inferior a la mitjana. Aquesta característica també ve donada pel coeficient d'asimetria que és positiu. El mínim d'aquesta variable és 18.72$ que fa pensar que es tracta d'una vivenda en la que els residents s'han traslladat fa poc i per tant no tenen un consum d'energia molt elevat. El màxim d'aquesta variable és 8121.56$ que es pot relacionar amb la despesa d'una familia adinerada o d'una familia molt nombrosa que viu en un habitatge gran amb molt consum. La curtosi és molt elevada el que indica que hi ha una quantitat molt elevada de dades que tenen aproximadament els mateixos valors o valors semblants.

Es realitza un histogràma per veure aquesta asimetria d'una forma més gràfica. A part, també es realitza un gràfic de caixa per detectar si hi ha presència d'outliers ja que tenint en compte l'elevat rang de la variable, és molt probable que n'hi hagi.

Respecte les variables independents, la mida mitjana de les llars és de 2384 peus quadrats que equivaldria a uns 221 metres quadrats. Els habitatges a Estats Units són molt més grans que els d'Espanya i aquestes xifres són normals. La variable que relaciona la superfície de la llar té una desviació típica molt elevada que, tenint en compte el valor semblant del coeficient d'asimetria amb la variable de la despesa d'energia, indica que en la base de dades hi ha cases molt grans. Això s'observa en el màxim que indica que hi ha una casa que té 8501 peus quadrats de superfície que equivaldria a 789.77 metres quadrats. La curtosi en aquesta variable també indica que hi ha un gran nombre de cases amb una superfície semblant però aquest apuntament és més lleuger que en la variable resposta.

La superfície considerada no finalitzada o no part de l'habitatge, hi ha diversos casos en el que les observacions tenen els mateixos valors per aquesta variable que per a la variable que descriu la superfície total de la casa per tant porta a pensar que hi ha diverses vivendes de la mostra que són àtics o habitatges encara en construcció. La desviació típica igual que en la superficie total és molt elevada però no hi ha tanta asimetria. Parlant de la variable que mesura la superfície de la casa que té calefacció, el comportament és bastant similar al de la variable superfície total ja que la major part de les llars tenen calefacció a tota la casa. El valor de l¡asimetria és exactament el mateix que en la superfície total. La desviació típica també és molt elevada de tal manera que hi haurà valors molt elevats en aquesta variable. La curtosis és també molt elevada en les dues variables per tant això indica que hi ha un gran nombre d'observacions similars.

Pel què fa al nombre de dormitoris, la mitjana es troba al voltant de 3 per casa. Les cases més grans obviament seran les que més habitacions tinguin arribant fins a un màxim de 10 dormitoris. En aquesta variable la desviació típica és molt baixa degut tabé al possible rang de les dades en el que no pot ser que hi hagi cases amb nombres molt elevats d'habitacions. En aquesta variable a diferència de les anteriors el coeficient d'asimetria és baix per tant les dades són bastant simètriques en aquesta variable però la curtosi segueix sent elevada, la qual cosa indicant apuntament.

El mateix passa amb el nombre de televisors. En mitjana n'hi ha entre 2 i 3 per casa. Igual que abans la desviació típica és baixa ja que el rang de les dades també és bastant limitat. Hi ha cases que no disposen de cap televisor mentre que hi ha alguna familia que disposa de fins a 9 televisors. El coeficient d'asimetria és baix i la curtosi molt elevada. Tot i que es podria pensar que hi ha un comportament semblant en el cas del nombre de ventiladors ja que ambdues variables indiquen la cantitat d'algun electrodomèstic, no és així.

En el cas del nombre de ventiladors, la mitjana es troba per sota d'un i mirant la mediana s'observa que la meitat de les cases no disposen de cap ventilador. Hi ha una asimetria molt elevada ja que hi ha cases que disposen de molts ventiladors. Aquesta és la variable amb més curtosi de les que s'ha utilitzat i té sentit que el valor sigui molt elevat ja que com a mínim el 50% de les dades té el mateix nombre de ventiladors que és 0.

El nombre de smartphones per casa en mitjana és de 1.68 fet que té bastant sentit si s'observa la mitjana del nombre de persones visquent a la casa que és una mica superior a 2 per tant aproximadament cada habitant té un smartphone tenint en compte que els nens no en solen tenir. La desviació típica de les dues variables també és baixa també degut al possible rang de les dades. Pel que fa a la asimetria hi ha certa asimetria per la dreta que indica valors bastant elevats. Les curtosis son semblants i indiquen cert apuntament en les variables.

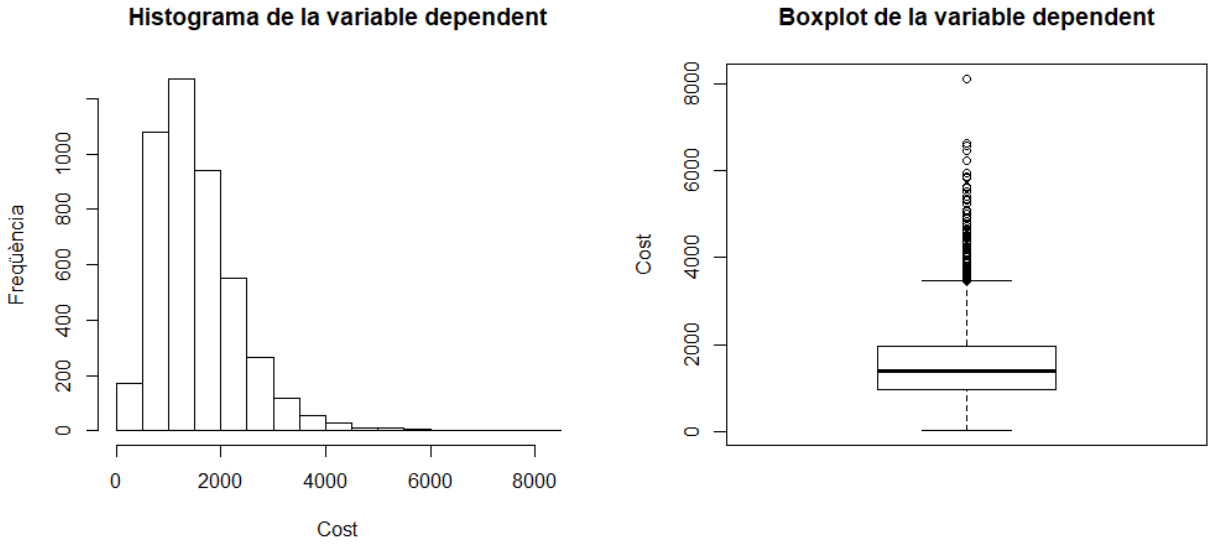


**Figura 3:** Gràfics descriptius de la variable dependent consum anual d'energia elèctrica en dòlars

Tal i com s'ha comentat amb anterioritat, s'han realitzat diversos gràfics per a la variable resposta que es poden veure en la Figura 3. A l'histograma s'aprecia una asimetria per la dreta molt elevada que indica valors molt elevats de la variable, fet que també es veu representat en el boxplot on es detecta un gran nombre d'outliers. Aquest fet és bastant lògic ja que la majoria de la gent té cases amb característiques bastant similars i per tant hi haurà un consum d'energia també similar. En canvi,la gent adinerada que representa una petita part de la població, tindrà cases més grans on el consum d'energia serà major i possiblement seran els que han estat identificats com a outliers. A continuació es dibuixa un gràfic per veure el creixement de la variable dependent en els seus quantils.

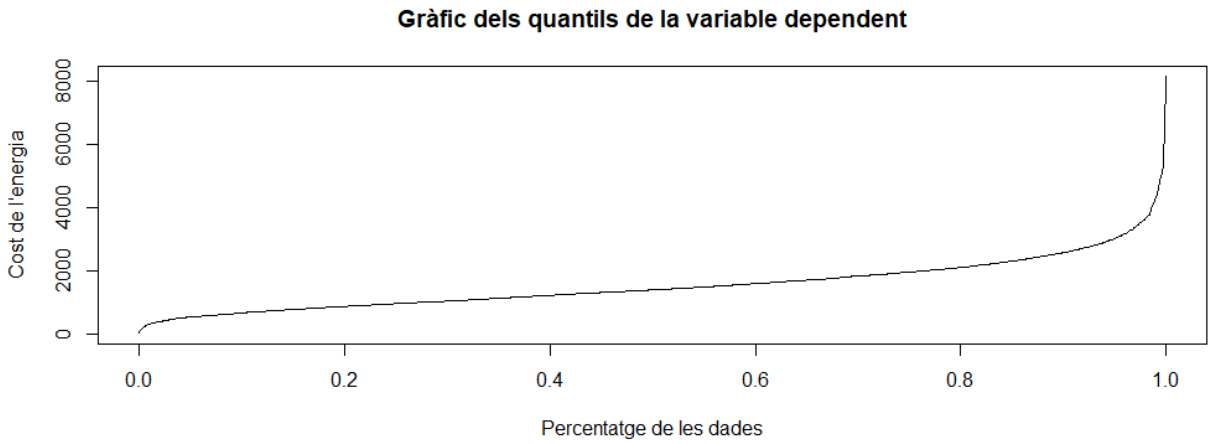


**Figura 4:** Gràfics dels quantils de la variable dependent consum anual d'energia elèctrica

Si es mira la distribució de la variable resposta en els seus quantils en el grafic de la Figura 4, es pot observar que el creixement és constant a mesura que creix el quantil però per als quantils a partir del quantil 90 els valors de la variable aumenten dràsticament.

## 5.2 Assegurances

La segona base de dades té relació amb la distància que recorren els conductors en excés de velocitat. És conegut que algunes entitats d'assegurances, per tal de posar un preu quan es demana una assegurança, demanen que durant un temps, el conductor, afegeixi un petit aparell al costat de la bateria del cotxe. Aquesta pràctica, tot i que encara no està gaire extesa, sembla que abara essent dominant. Aquest aparell recull dades bàsiques de conducció tals com la velocitat o el kilometratge. Aquesta base de dades ha estat distribuida per una assegurança de cotxes de la qual s'omet la identitat per motius de privacitat. La base de dades original disposa de 9614 observacions i 19 variables. Per a realitzar aquest treball s'ha agafat una submostra de 7691 conductors degut també a motius de privacitat i veient prèviament que la magnitud dels resultats no canviava en gran mesura. Respecte les variables, de la mateixa manera que en la base de dades anterior, no s'han utilitzat totes les variables disponibles sinó que s'ha fet una selecció d'algunes. A continuació, a la Taula 4, es pot trobar quines han estat les variables seleccionades per al treball i la seva descripció.

| Tipus | Variable | Descripció |
|---|---|---|
| Dependent | Toler_km_2010 | Nombre de kilometres conduïts per sobre del límit de velocitat. |
| Independent | lnKm | Logarítme del nombre de kilometres conduïits durant el 2010. |
| | Porc_vurba | Percentatge de kilometres conduïts en carreteres urbanes. |
| | Porc_nocturn | Percentatge de kilometres conduïts de nit |
| | Edad | Edat a l'inici del 2010. |
| | Sexo | Sexe (1 = Home, 0 = Dona). |

**Taula 4:** Descripció de les variables de la base de dades d'assegurances.

Per començar a treballar amb la segona base de dades, es comença fent estadística descriptiva sobre les diverses variables per tal de detectar valors anomals i veure les característiques que tenen les diferents variables seleccionades.

| Variable | Mitjana | Desviació Típica | Mediana | Mínim | Màxim | Asimetria | Curtosi |
|---|---|---|---|---|---|---|---|
| Toler_km_2010 | 1400.12 | 2008.67 | 689.42 | 0 | 23500.19 | 3.70 | 23.75 |
| lnKm | 9.26 | 0.76 | 9.37 | -0.37 | 10.96 | -1.95 | 13.38 |
| Porc_vurba | 26.36 | 14.29 | 23.47 | 0 | 100 | 1.56 | 6.33 |
| Porc_nocturn | 7.01 | 6.1 | 5.3 | 0 | 46.34 | 1.03 | 4.16 |
| Edad | 24.77 | 2.82 | 24.61 | 18.11 | 31.56 | 0.10 | 2.23 |

**Taula 5:** Estadístics descriptius de la segona base de dades

A part de les variables de la taula anterior, un 50.88% de les observacions són homes és a dir, hi ha aproximadament el mateix nombre d'homes que de dones. En relació amb la variable dependent, com s'observa a la Taula 5, la gent no sol conduir per sobre del límit de velocitat durant molts kilometres. Això es sap perquè la mitjana és molt superior a la mediana i hi ha molts casos en que el conductor no ha superat en cap moment el límit de velocitat. Aquest fet també pot ser comprovat mirant el coeficient d'asimetria que és positiu i molt elevat. La curtosi és molt alta degut a què la major part de la gent no supera el límit de velocitat i si ho fan, ho fan durant molt pocs kilometres. A la taula, s'observa també que hi ha valors molt elevats ja que el valor més alt és de 23500.19 kilometres per sobre del límit de velocitat. És probable que aquests valors elevats estiguin relacionats amb l'edat dels conductors ja que es creu que quan més jove és el conductor més es condueix per sobre els límits de velocitat. Una altra raó pot ser la professió del conductor, que requereixi molts desplaçaments i l'àmibt on condueix.

Els kilometres conduïts pels diversos conductors es troben en escala logarítmica. En mitjana, la gent recorre uns 10.000 kilometres en el periode durant el que tenen l'aparell de mesura que és d'un any i la

mediana és bastant similar. Tot i que a l'escala logarítmica sembla que no hi ha distàncies recorregudes molt elevades, el valor màxim ha recorregut 57.000 kilometres. El coeficient d'asimetria és negatiu i bastant elevat i indica que hi ha certa asimetria per l'esquerra, representant que hi ha més valors baixos que elevats, és a dir, en conjunt, certs conductors no agafen molt sovint el cotxe. La curtosi en aquesta variable també és molt elevada això implica que la major part de les dades tenen valors semblants.

Parlant de les zones por on solen conduir els conductors, aquests no ho solen fer per àrees urbanes ja que la mitjana es troba en que els conductors condueixen un 26% del temps per aquest tipus de carreteres. Sí que és cert que el valor màxim és 100 que indica que hi ha conductors que només condueixen per àrees urbanes, probablement només agafen el cotxe per anar de casa a la feina en zones metroploitanes. En aquesta variable també hi ha certa asimetria per la dreta i la curtosi també és elevada.

Fent referència a quan solen conduir els conductors de la mostra, aquests no solen conduir de nit ja que en mitjana ho fan només un 7% de la distància total. El valor més alt en aquest cas és un 46.34% que pot ser per dues raons; o bé el conductor treballa de nit o bé és una persona, possiblement jove, que utilitza el cotxe per oci. En aquest cas també hi ha certa asimetria per la dreta però és bastant lleugera. També hi ha certa curtosi però és menys forta que en les variables anteriors d'aquesta base de dades.

Per últim el rang d'edat dels conductors es situa entre els 18 i els 32 anys per tant és una mostra bastant jove i la mitjana de la mostra és de 24. El coeficient d'asimetria és molt baix per tant les observacions d'aquesta variable son simètriques i hi ha poca curtosi és a dir les dades estan bastant repartides entre els possibles valors. La raó que no hi hagi conductors grans en aquesta mostra prové del fet que l'entitat asseguradora només va comercialitzar aquesta assegurança a gent jove.

De la mateixa manera que en la base de dades del consum d'energia, s'estudia més a fons la variable dependent començant per un histograma i un gràfic de caixa.

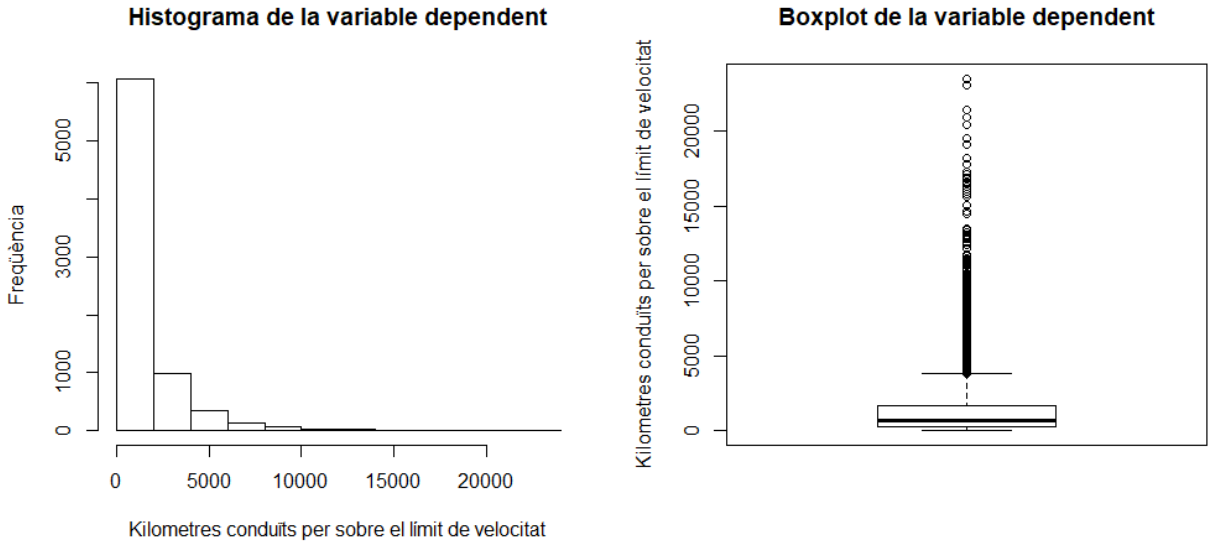


**Figura 5:** Gràfics descriptius de la variable dependent distància recorreguda per sobre del límit de velocitat permesa

En l'histograma de la Figura 5 s'observa que la major part dels conductors no superen el límit de velocitat durant gran part de la conducció però en alguns casos es supera en gran mesura. En el gràfic de caixa es pot veure aquest fet més clarament per la gran cantitat d'outliers que apareixen per la part superior de la caixa. És normal que els gràfics surtin d'aquesta manera ja que la gent sol respectar les normes de conducció. En alguns casos el límit de velocitat es pot superar per distraccions o per a realitzar algun avançament i per això la majoria de valors són baixos. Els valors elevats és molt probable que corresponguin a gent jove que en general té comportaments d'alt risc en la conducció i a gent que hagi recorregut molts kilometres ja que es creu que la distància total i la distància amb excés de velocitat són variables molt relacionades. Igual que en l'altre base de dades es dibuixa el gràfic que mostra els valors dels quantils de la variable resposta.

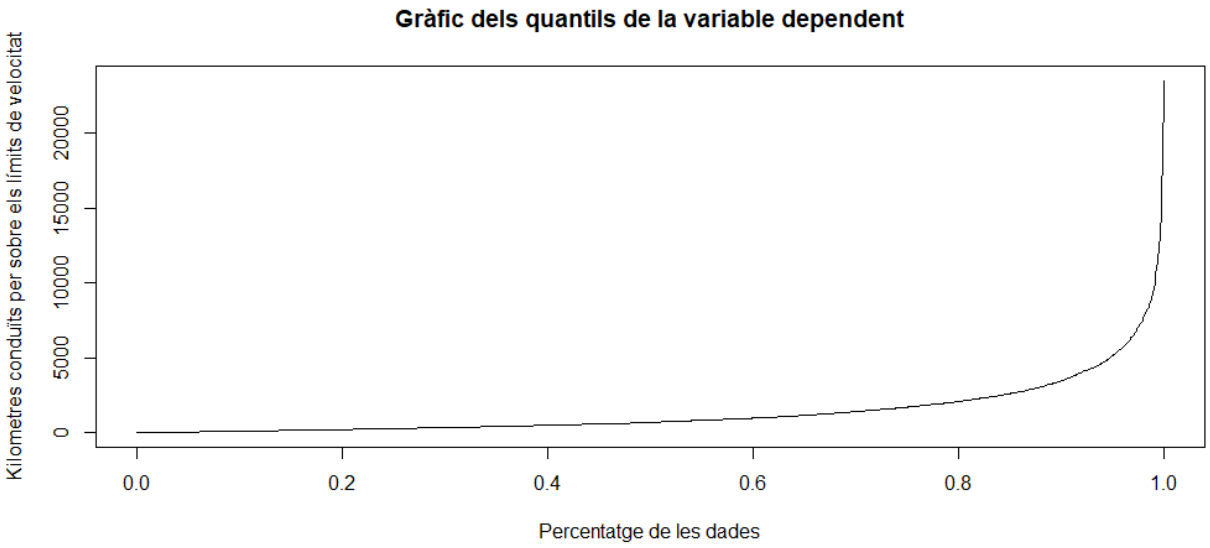


**Figura 6:** Gràfics dels quantils de la variable dependent distància recorreguda per sobre del límit de velocitat permesa

En el gràfic de la Figura 6 s'observa un comportament de la variable dependent bastant similar al de la base de dades del consum d'energia. El creixement del nombre de kilometres és pràcticament constant a mesura que va creixent el quantil però a partir del quantil 90 el creixement és molt més vertical i pot portar problemes a l'hora d'ajustar un model lineal.

# 6 Implementació

Per a la realització d'aquest treball s'ha utilitzat el software estadístic anomenat R i en aquest apartat es faran una sèrie de comentaris sobre les funcions que s'han utilitzat per a estimar la regressió quantílica tant per els quantils com per al TVaR.

## 6.1 La funció rq

El programa R té un paquet anomenat `quantreg` creat per Koenker i altres autors [17] que està enfocat a la regressió quantílica entre altres coses. Aquest paquet disposa d'una funció `rq` que ajusta la regressió quantílica i retorna els coeficients de l'efecte estimat d'un conjunt de variables juntament amb la desviació típica de l'estimació, el valor de l'estadístic t i el p-valor del contrast d'hipòtesi que estudia si el coeficient es pot considerar 0. La comanda que s'utilitza per ajustar dita regressió és:

```
rq(formula = , tau = , data =  , subset = , weights = , na.action = ,
   method = , model = , contrasts = , ...)
```

S'ha decidit no posar el codi sencer de la funció ja que aquest utilitza una sèrie de funcions que també han estat programades per Koenker i que estan incloses en el paquet `quantreg`. Tot i així sí que es farà una explicació dels paràmetres que s'han d'introduir a la funció per tal d'obtenir els resultats. A la publicació [16], es pot trobar una explicació bàsica sobre el funcionament d'aquesta funció i es plantegen diversos exemples sobre el seu ús.

En primer lloc hi ha el paràmetre `formula`. Aquest és un paràmetre fonamental ja que és on s'especifiquen les variables dependents i la variable resposta. Per tal d'especificar-ho cal fer-ho de la manera:
"Variable resposta $\sim$ Variable dependent 1 + Variable dependent 2 +..."així s'indica quines són les diferents variables per a ajustar el model.

El segon paràmetre és `tau` i de la mateixa manera que s'ha fet en aquest treball, indica sobre quin quantil es vol fer la regressió. És un paràmetre opcional ja que si no s'especifica cap valor, la regressió quantílica s'ajusta per defecte al quantil 0.5 o el que seria la mediana. El valor d'aquest, igual que per als quantils, ha d'estar entre 0 i 1. Si el valor d'aquest paràmetre no està entre 0 i 1, no surt cap missatge d'error sinó que la funció intenta fer l'ajust i en comptes que apareguin els resultats de l'ajust, apareixen els valors de la funció pèrdua del quantil per a cada observació.

El següents dos paràmetres de la funció indiquen aspectes de la base de dades que s'ha d'utilitzar. El paràmetre `data` és també de gran importancia per al model ja que és el que indica la base de dades de la que

s'han d'obtenir els valors de les variables. Per altra banda, el paràmetre `subset` és opcional i és un paràmetre que serveix per indicar si es vol que s'utilitzin unes observacions de la base de dades en concret en comptes de totes les observacions.

El següent paràmetre és el paràmetre `weights` que serveix per donar pesos a les observacions depenent de la importancia que tinguin. En aquest treball s'ha considerat que totes les variables tenien el mateix pes i per tant aquest paràmtre no s'ha fet servir. En cas de que les observacions tinguessin pesos la funció objectiu que caldria minimitzar canvia lleugerament. De fet, a la base de dades del consum d'energia, hi havia ponderacions mostrals però per a aquest treball no s'han fet servir i es deixa aquesta qüestió per a una possible millora futura.

El paràmetre `na.action` és un paràmetre opcional que només és útil quan hi ha NAs o dades faltants a les observacions. Si hi han NAs a la base de dades, la funció *rq* no aconsegueix ajustar un model a les dades i surt un missatge d'error. El que es sol fer normalment quan hi ha observacions que tenen NAs és utilitzar la comanda `na.omit` que els elimina.

El següent paràmetre és el paràmetre `method` que serveix per especificar quin és el mètode que es vol fer servir per ajustar el model. Per defecte el mètode que s'utilitza és *"br"* que fa referència al mètode de Barrodale i Roberts aquest mètode serveix per bases de dades amb uns pocs milers d'observacions. Altres mètodes que s'accepten són *"fn"* que fa referència al mètode Frisch and Newton per a bases de dades més grans, *"pfn"* que és una variant del mètode Frisch and Newton que s'utilitza per a bases de dades encara més grans, *"sfn"* que és una altra variant del mètode de Frisch and Newton per a bases de dades amb moltes variables i amb moltes variables que siguin factors, *"fnc"* que permet especificar condicions en els paràmetres del model com per exemple que $\beta_1 \geq 1$ i per últim estàn els mètodes *l·lasso"* i *"scad"* que apliquen penalitzacions a les observacions per tal d'ajustar el model.

El paràmetre `model` serveix per si després d'ajustar el model de regressió quantílica es vol fer un summary d'aquest. Si no s'especifica el contrari després de fer l'ajust es deixarà el model ajustat en un format per fer el summary d'aquest.

Per últim hi ha el paràmetre `contrasts` que serveix per especificar, si a part de fer el contrast d'hipòtesis sobre els coeficients que s'ha explicat anteriorment, es vol estudiar algun altre contrast en concret. Els punts suspensius fan referència a què poden haver-hi paràmetres addicionals per a la funció que provinguin de les funcions `rq.fit.br` i `rq.fit.fnb`, ja que ambdues funcions es troben dins del paquet quantreg.

## 6.2 Ús de la funció optim per a la regressió quantílica

En aquesta part es presenta la funció que s'ha programat per tal d'obtenir els ajustos de la regressió quantílica per als quantils $\tau = (0.25,\ 0.5,\ 0.75$ i $0.9)$ a diferència de la funció anterior aquí sí que es mostra tot el codi de les diverses funcions que s'han programat i es fa una explicació sobre com funcionen.

```
coeficientes <- function(params){
  aux <- Y - X%*%params
  mult1 <- ifelse(aux>=0,quant,(quant-1))
  t(mult1)%*%aux
  }
```

La primera funció és la funció `coeficientes` i l'únic que fa és calcular el valor de la funció objectiu d'una formula utilitzant les variables de la base de dades i els coeficients estimats. `aux` serveix per a calcular la funció de pèrdua del quantil i correspon a la condició de la funció indicadora. `mult1` és el valor de la funció pèrdua que serà $\tau$ si `aux` és positiu i $\tau - 1$ si és negatiu. per últim, el producte de `mult1` i `aux` és el que dona el valor de la funció objectiu. Aquesta funció és necessaria per tal d'utilitzar la funció `optim`. La funció `optim` el que fa és minimitzar el valor d'una funció objectiu (que en el cas d'aquest treball serà coeficientes) canviant una sèrie de variables que s'han d'especificar (en aquest cas seran les $\beta$).

```
optim(betaINI,coeficientes,control=list(maxit=2000000, beta = 0.01))
```

La funció `optim` s'ha utilitzat tal i com es mostra a sobre. BetaINI correspon als valors que s'han donat als coeficients de forma provisional per tal de que `optim` comenci a buscar quins valors de $\beta$ minimitzen la funció objectiu. Aquests valors inicials s'explicarà quins són i perquè s'han triat a la secció 7. El segon paràmetre de la funció és la funció objectiu a minimitzar que com s'ha dit és `coeficientes`. Per últim, el paràmetre control és un conjunt d'especificacions per al mètode d'optimització que també s'explicarà com s'han triat a la secció 7 d'aquest treball. A continuació es presenta la funció que s'ha utilitzat per fer el remostreig de les observacions i calcular els coeficients.

```
standardboot <- function(Y,X,n){


  beta90 <- betas
  mostra <- sample(x = 1:length(Y), size = n, replace = FALSE)
  Yaux <- Y[mostra]
  Xaux <- X[mostra,]
```

```
  coeficientes90 <- function(params){
    aux <- Yaux - Xaux%*%params
    mult1 <- ifelse(aux>=0,quant,(quant-1))
    t(mult1)%*%aux
  }

  betaaux <- optim(beta90,coeficientes90,control=list(maxit=200000, beta = 0.01))
  beta90 <- betaaux$par

  beta90

}
```

Per aquesta funció és necessari aportar la variable resposta que dins la funció serà Y, una matriu amb les variables explicatives que dins la funció serà X i un valor n que correspon al nombre d'observacions que es vol obtenir en el remostreig. El primer que fa aquesta funció és generar n nombres aleatoris sense reposició entre 1 i el nombre d'observacions i guarda en un vector i una matriu (Yaux i Xaux respectivament) les observacions corresponents als nombres aleatoris escollits. A continuació es torna a definir la funció a optimitzar però aquesta vegada utilitzant `Yaux` i `Xaux` per calcular el valor de la funció objectiu. per últim, s'utilitza la funció `optim` per a obtenir l'estimació dels efectes del model ajustat. Utilitzant el codi següent es repeteix aquesta funció 1000 vegades i els guarden els valors dels coeficients en una matriu.

```
tini <- Sys.time()
for(i in 1:nrep){
  coefboot[i,] <- standardboot(Y,X, n)
  print(paste0("Repetició ", i, " de 1000."))
}
tfin <- Sys.time()

(tfin - tini)
```

La funció `Sys.time` serveix per guardar la hora que marca l'ordinador que s'executa la comanda. Es guarda a `tini` l'hora en que es comença a executar les repliques de la funció anterior i a `tfin` l'hora que acaben d'executarse. La resta `tfin - tini` indica quina ha estat la durada del procediment. A més a més, aquesta funció imprimirà un missatge per pantalla que indicarà per quina replica va. Tant la funció standardboot com aquesta són molt llargues d'executar i en continuacions d'aquest treball es podria intentar

optimitzar-les per tal d'obtenir els resultats d'una forma més ràpida. Un cop s'hagin guardat tots els valors dels coeficients, s'executen les següents línies de codi.

```
desv <- rep(NA,11)
for(i in 1:11){
  desv[i] <- sd(coefboot[,i],na.rm = TRUE)
}

t.value <- (coeficients-0)/desv
pval <- ifelse(t.value > 0,2*(1 - pnorm(t.value)),2*pnorm(t.value))
resultats <- data.frame("Coefficients" = coeficients ,"sd" = desv,"t.value" = t.value,
                        "p.value" = pval)
row.names(resultats) <- noms
resultats
```

En primer lloc, el bucle que es troba al principi d'aquest últim codi correspon al càlcul de les desviacions típiques de les estimacions. El que fa és calcular la desviació típica per a cada columna de la matriu. `T.value` es defineix com al valor de l'estadístic t que s'ha comentat en la secció 3.3 i el `pval` és el p-valor de l'estadístic t. Els resultats surten en forma de taula on les columnes correspondran al valor de l'efecte estimat, la seva desviació típica, l'estadístic t del contrast d'hipòtesis i al p-valor de cada variable.

## 6.3 Ús de la funció óptim per al cas inspirat en el TVaR

El codi i les funcions utilitzats per a estimar el TVaR són bastant similars als utilitzats amb la funció `optim` per a la regressió quantílica a un quantil. De la mateixa manera que abans es presentarà el codi complet per a aquesta part del treball.

```
coefTVaR <- function(params){

  aux <- Y - X%*%params
  ind <- ifelse(aux>=0,0,1)
  mult2 <- 1/3 - ind/2 - quant^3/3 + (ind*(quant^2))/2
  t(aux)%*%mult2

}
```

En primer lloc es necessita definir la funció objectiu que es volia minimitzar. En aquest cas i tal i com s'ha vist a la secció 4.2 la funció de pèrdua del quantil era:

$$\rho(\tau) = \left(\frac{1}{3} - \frac{I((Y - X\beta) \leq 0)}{2} - \frac{\tau^3}{3} + \frac{\tau^2 I((Y - X\beta) \leq 0)}{2}\right) \tag{17}$$

Es defineix `aux` com la condició de la funció indicadora i `ind` com el valor 0 o 1 d'aquesta funció. Igual que en la funció per a la regressió quantílica mult2 és la funció pèrdua que té la fòrmula escrita en l'equació (17). De la mateixa manera que abans aquesta definició de la funció objectiu serveix per a poder utilitzar la funció `optim` per trobar els valors òptims dels coeficients. L'altre punt del codi on hi ha diferències entre la regressió quantílica per al quantil i per al TVaR és en la funció `standardboot`.

```
standardboot2 <- function(Y,X,n){

  valors <- betasTVAR
  mostra <- sample(x = 1:length(Y), size = n, replace = FALSE)
  Yaux <- Y[mostra]
  Xaux <- X[mostra,]

  coefsdTVaR <- function(params){

    aux <- Yaux - Xaux%*%params
    ind <- ifelse(aux>=0,0,1)
    mult2 <- 1/3 - ind/2 - quant^3/3 + (ind*(quant^2))/2
    t(aux)%*%mult2

  }

  betaaux <- optim(valors,coefsdTVaR,control=list(maxit=200000, beta = 0.01))
  valors <- betaaux$par

  valors
}
```

En aquesta funció també es requereix la provisió d'una variable resposta Y, una matriu amb les variables descriptives X i la quantitat d'observacions que es volen agafar per a realitzar el remostreig. Es tornen a

seleccionar els nombres aleatoris sense reposició per tal de seleccionar les observacions de forma aleatòria i generar les submostres per calcular les desviacions típiques de les estimacions. A part d'aquestes dues funcions, l'obtenció dels resultats en format de taula, el càlcul de les desviacions típiques i l'execució de 1000 vegades d'aquesta funció es fa de la mateixa manera que en la regressió quantílica per a un quantil.

# 7 Resultats

Anteriorment s'ha vist quina era la fòrmula per calcular els coeficients de la regressió quantílica i com es pot estimar la desviació típica d'aquests coeficients utilitzant el mètode bootstrap. El primer que s'ha fet és, comparar els resultats de l'ajust de la regressió quantílica al quantil 90 entre les dues fucions, la creada i la predefinida, per tal de veure si està ben programada.

## 7.1 Consum d'energia

Hi ha diversos mètodes per a realitzar la optimització dins la funció `optim` però cal valorar quin ens dona el millor resultat quan es resol el problema d'optimització. Es planteja escollir el mètode de la funció `optim` en funció del valor obtingut a la funció objectiu i el temps de computació ja que aquest serà fonamental a l'hora d'aplicar el mètode bootstrap. Els valors dels coeficients inicials són els obtinguts utilitzant el model lineal que seran presentats més endavant. A la taula següent es poden veure els resultats dels diferents mètodes d'optimització.

| Mètode | Temps | Funció objectiu |
|---|---|---|
| BFGS | 0.21 segons | 636825.4 |
| Nelder-Mead | 10.4 segons | 636501.5 |
| Conjugate Gradients | 0.32 segons | 668057.6 |
| L-BFGS-B | 0.62 segons | 668062.1 |

**Taula 6:** Valors de la funció objectiu i temps de computació dels mètodes d'optimització

Com es pot observar a la Taula 6, el mètode BFGS és el més ràpid i ens dona un bon resultat a la funció objectiu però, el mètode Nelder-Mead triga bastant més però redueix bastant el valor de la funció objectiu aleshores cal valorar quin d'aquests dos mètodes s'utilitzarà per realitzar l'optimització. El mètode escollit en aquest treball ha estat el de Nelder-Mead ja que s'ha prioritzat el valor de la funció objectiu sobre el temps de computació. Els mètodes de *Conjugate Gradients* i *L-BFGS-B* tenen resultats molt semblants entre ells, el temps de computació és baix però el valor de la funció objectiu és molt més elevat que en els altres dos casos i això fa que siguin descartats directament. L'únic problema d'haver escollit el mètode de Nelder-Mead per a realitzar l'optimització és que quan es faci el mètode bootstrap per a calcular les desviacions típiques, el temps de computació serà molt elevat.

Durant la realització d'aquest treball es va arribar a un resultat molt important el qual és necessari comentar. Un cop seleccionat el mètode d'optimització va caldre determinar quins haurien de ser els valors inicials dels coeficients. Es va provar a realitzar l'optimització desde diversos punts inicials i es va veure que en tots ells la funció convergia però els valors de la funció objectiu eren diferents cada vegada. Es presenten

els resultats de les funcions objectius partint desde diversos punts.

| | Funció objectiu |
|---|---|
| Partint de $\beta = 0$ | 887175.9 |
| Partint de $\beta = 1$ | 814031.0 |
| Partint de $\beta = 500$ | 1881703.0 |
| Partint de $\beta = \beta_{lm}$ | 636501.5 |
| Partint de $\beta = 1000$ | 3997845.0 |

**Taula 7:** Valors de la funció objectiu optimitzant desde diversos punts inicials

Com s'observa a la Taula 7, partint desde diversos punts a l'hora d'optimitzar, s'obtenen valors de la funció objectiu diferents i per tant els valors dels coeficients estimats també canvien. Això vol dir que la funció que s'està optimitzant té mínims locals i és fàcil que el mètode d'optimització convergeixi a un d'ells quan el que es desitja és que el mètode convergeixi al mínim absolut. La quantita de mínims locals que hi ha a la funció és bastant elevada ja que si es comença fent que totes les $\beta$ siguin iguals a 0 ja s'obté un resultat diferent que fent que totes les $\beta$ siguin iguals a 1. Una continuació d'aquest treball sería explorar en detall els diferents mètodes d'optimització i tunejar els paràmetres per intentar arribar al mínim absolut. En vista dels valors de la taula, es decideix realitzar primer una estimació per mínims quadrats ordinaris i prendre els valors dels coeficients d'aquesta estimació com a valors inicials ja que són els que donen un valor de la funció objectiu més baix. També es veu que si es prenen valors de $\beta$ inicials molt grans, el valor de la funció objectiu és molt elevat.

Un cop establerts tant el punt desde el qual es comença a calcular l'optimització com el mètode per fer-ho, es passa a calcular els valors dels coeficients de la regressió quantílica per a diversos quantils. Aquests es presenten a continuació en forma de taula.

| Variable | $\beta_{lm}$ | $\beta_{rq0.25}$ | $\beta_{op0.25}$ | $\beta_{rq0.50}$ | $\beta_{op0.50}$ | $\beta_{rq0.75}$ | $\beta_{op0.75}$ | $\beta_{rq0.90}$ | $\beta_{op0.90}$ |
|---|---|---|---|---|---|---|---|---|---|
| Constant | 512.22 | 277.19 | 94.03 | 478.84 | 339.19 | 697.07 | 514.69 | 939.81 | 843.49 |
| TOTSQFT_EN | 0.031 | 0.005 | 0.006 | 0.043 | 0.040 | 0.012 | 0.012 | 0.066 | 0.065 |
| SWIMPOOL | 650.24 | 486.27 | 485.52 | 569.45 | 582.50 | 761.62 | 766.36 | 866.54 | 868.75 |
| TOTCSQFT | 0.12 | 0.12 | 0.12 | 0.14 | 0.14 | 0.12 | 0.12 | 0.10 | 0.103 |
| TOTHSQFT | -0.023 | -0.033 | -0.035 | -0.089 | -0.085 | -0.018 | -0.019 | 0.04 | 0.043 |
| BEDROOMS | 66.14 | 50.83 | 51.18 | 79.29 | 79.51 | 84.26 | 83.77 | 68.89 | 71.30 |
| SOLAR | 179.07 | 183.03 | 182.19 | 125.80 | 133.24 | 185.13 | 185.66 | 37.84 | 90.97 |
| TVCOLOR | 106.04 | 92.54 | 92.10 | 101.93 | 102.65 | 128.23 | 128.17 | 141.81 | 141.50 |
| NUMSMPHONE | 25.61 | 16.25 | 16.51 | 35.03 | 33.83 | 24.20 | 25.36 | -3.90 | -3.68 |
| NUMFLOORFAN | -1.85 | 2.52 | 2.46 | -9.40 | -7.56 | 5.47 | 5.31 | 8.20 | 8.92 |
| NHSLDMEM | 76.93 | 56.98 | 57.18 | 68.27 | 68.63 | 110.74 | 110.14 | 152.46 | 151.70 |
| Funció Objectiu | - | 887667.1 | 837568.3 | 1196495 | 1172916 | 1073849 | 1044440 | 637309.6 | 636501.5 |

**Taula 8:** Efectes de les variables dependents en la regressió quantílica per a diferents quantils.

Com es pot observar a la Taula 8 els valors de les funcions objectiu utilitzant la funció `optim` són més baixes que els de la funció `rq` on sempre es troben diferències en les estimacions de la constant i en el cas de la regressió pel quantil 0.9 també hi ha diferència en l'estimació de la variable SOLAR. És interessant afegir els valors dels coeficients realitzant la regressió lineal perquè com per exemple en el cas de la variable TOTCSQFT els valors dels coeficients de les regressions per als diferents quantils són sempre semblants als de la regressió lineal. No s'indica el valor de la funció objectiu del model lineal perquè el problema d'optimització que es resol per calcular els coeficients és diferent que el de la regressió quantílica. Per tal de veure l'evolució dels efectes de les variables amb més claredat es presenta un gràfic que representa els efectes dels coeficients a mesura que es va augmentant el quantil.

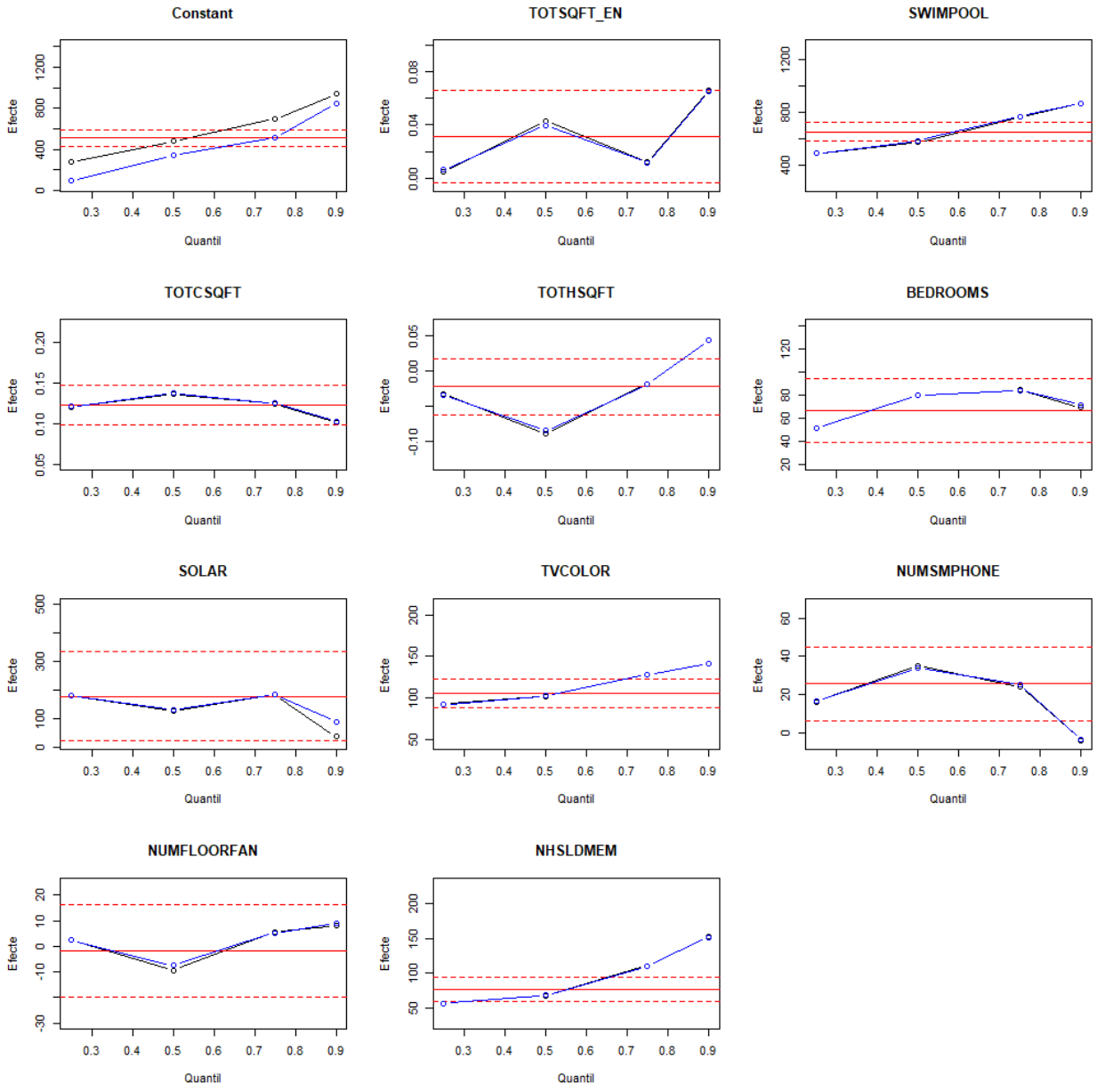


**Figura 7:** Gràfics dels efectes de les variables a mesura que augmenta el quantil en la base de dades de consum d'energia.

En els gràfics de la Figura 7, la línia negra correspon als valors del coeficients estimats utilitzant `rq` per a la regressió quantílica. La línia blava, correspon als coeficients de la regressió quantílica utilitzant la funció creada amb `optim`. La línia vermella continua representa els coeficients de la regressió lineal i les dues línies discontinues que l'envolten representen els intervals de confiança d'aquests coeficients.

Si tots els valors dels efectes de la regressió quantílica de cada variable es troben dins l'interval de la regressió lineal, es considera que l'efecte de la regressió quantílica per aquella variable és igual a l'efecte de la regressió lineal. Es veu que en alguns casos, hi ha algunes variables com per exemple la variable BEDROOMS i la variable NUMFLOORFAN que tenen tots els valors per als diferents quantils dins l'intèrval per tant els dos tipus de regressió estimen el mateix efecte.

L'evolució de les diferents variables a mesura que augmenta el quantil es molt semblant comparant les dues funcions per a estimar la regressió quantílica. Tal i com s'ha comentat, la major diferència es troba en l'estimació de la constant i en l'últim quantil amb la variable SOLAR mentre que per les altres variables els efectes estimats són molt semblants.

En el cas de la constant es veu que pels quantils més baixos, el valor és més baix i que a mesura que creix en quantil també augmenta l'efecte constant això i que excepte per al quantil 0.5 en el cas de la funció `rq` i per al quantil 0.75 en el cas de la funció `optim` els valors dels efectes es troben fora de l'interval de confiança de l'efecte del model lineal.

Per a la variable TOTSQFT_EN, a mesura que el quantil augmenta, l'efecte no té per què fer-ho. El valor del quantil 0.75 és més baix que el del 0.5 i després torna a creixer. En aquest cas les dues funcions estimen valors semblants per als efectes però en cap cas es surt de dins l'interval de confiança del model lineal per tant l'efecte pot considerar-se igual al del model lineal.

Per a la variable SWIMPOOL les estimacions es comporten bastant semblant a l'efecte de la variable constant. En aquest cas els valors de les dues funcions programades en R tenen valors molt semblant i per als primers quantils i els últims els coeficients es troben fora de l'interval de confiança del model lineal.

Fent referència a la variable a la variable TOTCSQFT les dues funcions estimen els mateixos efectes. Del quantil 0.25 al 0.5 l'efecte creix però del 0.5 cap endevant l'efecte de la variable decreix però mai surt de l'interval de confiança del model lineal per tant també es pot considerar que aquesta variable en la regressió quantílica té el mateix efecte que el model lineal.

Per a la variable TOTHSQFT, per al quantil 0.25 l'efecte es considera igual al del model lineal i decreix arribant a un valor negatiu al quantil 0.5 i sortint de l'interval de confiança. A partir del quantil 0.5 l'efecte torna a creixer i per al quantil 0.9 l'efecte també es considera diferent al del model lineal.

En quant al nombre de dormitoris l'efecte comença creixent des del quantil 0.25 fins al 0.75 i després decreix lleugerament. Per a tots els quantils les funcions estimen els mateixos efectes i en cap moment es pot considerar diferent que l'efecte del model lineal.

Per a la variable SOLAR, l'efecte de la variable comença decreixent fins al quantil 0.5, del quantil 0.5 al 0.75 torna a creixer i del 0.75 al 0.9 torna a decreixer. Per al quantil 0.9 les estimacions de les dues funcions són diferents però en cap dels dos casos es surt de l'interval de confiança del model lineal. Els creixements i decreixements de l'efecte, tot i ser elevats, no surten de l'interval de confiança de l'efecte del model lineal degut a l'elevada desviació típica d'aquest.

Per al nombre de televisors a color l'efecte creix del quantil 0.25 al 0.9. Els efectes estimats de les dues funcions són molt semblants i per al quantil 0.25 l'efecte es considera igual al del model lineal i a partir del 0.75 ja es pot considerar diferent.

En respecte al nombre de telèfons mòbils, el comportament de l'efecte de la variable és el mateix que en la variable TOTCSQFT. Comença amb un creixement important des del quantil 0.25 fins al quantil 0.5. A partir del quantil 0.5 comença a haver-hi un decreixement lleu que es pronuncia a partir del quantil 0.75 arribant a sortir de l'interval de confiança en el quantil 0.90.

Una altre variable en la que no es diferencia l'efecte entre la regressió lineal i la regressió quantílica és el nombre de ventiladors a la casa. L'efecte comença decreixent i a partir del quantil 0.5 creix fins al quantil 0.9. Sembla que per al quantil 0.5 hi ha certa diferencia entre la funcio `optim` i la funció `rq` on la primera dona un valor lleugerament més elevat.

Per últim, el nombre d'habitants de la casa sí que mostra diferències entre l'efecte estimat per la regressió lineal i la regressió quantílica. L'efecte creix desdel primer moment però no és fins al quantil 0.75 que es pot considerar que els efectes són diferents als de la regressió lineal. Les funcions estimen valors molt semblants per a tots els quantils.

Tenint en compte la taula dels efectes i els gràfics que s'acaben de comentar, s'observa que en una gran part de les variables l'efecte de la regressió lineal i el de la regressió quantílica es poden considerar iguals però,

com hi ha unes quantes variables en la que això no es compleix, és convenient aplicar la regressió quantílica si realment es vol explicar com afecta el valor d'un regressor en el quantil de la resposta. També es veu que en la major part dels casos en els que es considera que hi ha diferències entre els dos tipus de regressió, sol ser en els quantils més elevats. Degut a aquesta diferència d'efectes entre regressions per al quantil 0.9 i com que un dels objectius d'aquest treball és estudiar una mesura del risc, es decideix centrar-se en els efectes del quantil 0.9.

| Variable | $\beta_{rq}$ | $\beta_{optim}$ | $\text{Sd}_{rq}$ | $\text{Sd}_{optim}$ | $\text{p.valor}_{rq}$ | $\text{p.valor}_{optim}$ |
|---|---|---|---|---|---|---|
| Constant | 939.81 | 843.49 | 83.69 | 119.81 | 0 | 0 |
| TOTSQFT_EN | 0.066 | 0.065 | 0.039 | 0.018 | 0.091 | 0.0003 |
| SWIMPOOL | 866.54 | 868.75 | 114.35 | 56.53 | 0 | 0 |
| TOTCSQFT | 0.1 | 0.103 | 0.033 | 0.017 | 0.002 | 0 |
| TOTHSQFT | 0.043 | 0.043 | 0.04 | 0.025 | 0.27 | 0.09 |
| BEDROOMS | 68.88 | 71.30 | 30.12 | 19.94 | 0.022 | 0 |
| SOLAR | 37.84 | 90.97 | 124.16 | 106.44 | 0.761 | 0.39 |
| TVCOLOR | 141.81 | 141.50 | 18.35 | 10.05 | 0 | 0 |
| NUMSMPHONE | -3.9 | -3.68 | 21.94 | 11.54 | 0.86 | 0.75 |
| NUMFLOORFAN | 8.2 | 8.92 | 24.02 | 9.92 | 0.73 | 0.37 |
| NHSLDMEM | 152.46 | 151.70 | 22.13 | 10.86 | 0 | 0 |

**Taula 9:** Regressió quantílica al quantil 0.9 de les dades de consum d'energía

Com s'observa a la Taula 11, els valors de $\beta$ obtinguts a partir de la funció que s'ha creat són bastant similars als de la funció `rq`. La constant és un dels pocs casos on la funció creada i la funció `rq` difereixen sobre l'efecte. La funció `optim` estima un valor constant més baix però estima un valor de la desviació típica més elevat que la funció `rq`. El coeficient indica que per al quantil 0.9 hi ha un consum d'energia base de 843.49$. En cap de les dues funcions es considera que l'efecte de la constant no sigui important per al model. Totes les interpretacions que es faran suposaran que la resta de variables es mantenen igual.

Respecte la variable TOTSQFT_EN, és una variable a la que s'estima un efecte molt baix en les dues funcions. Això implica valors més baixos en l'estimació de la desviació típica. En aquest cas els efectes estimats entre les dues funcions són pràcticament idèntics però la desviació típica és més baixa en la funció `optim`. Tot i que el valor del coeficient és molt proper a 0, al tenir una desviació típica molt baixa, el p-valor per a la funció `optim` indica que és una variable important per al model. Com que la desviació típica estimat per la funció `rq` és més elevat fa que el p-valor sigui 0.09 així que depenent de la rigorositat que es vulgui tenir es pot considerar que la variable té efecte. El valor del coeficient indica que per cada peu quadrat que tingui la casa de superfície, el quantil 0.9 del cost en energia augmenta 0.065$. És molt rellevant assenyalar que al model lineal s'obtenia un valor del coeficient igual a 0.031 que en aquest cas indíca que un peu quadrat implica un increment mitjà del cost en energia de 0.031$

La variable SWIMPOOL és una altra variable en la què s'estima un coeficient molt elevat. Els coeficients entre les dues funcions són molt semblants i indiquen que el fet de tenir piscina fa augmentar al quantil 0.9 del consum d'energia en 868.75$. De la mateixa manera que en la variable anterior, la desviació típica estimada per la funció `optim` és molt més baixa que en la funció `rq` però en aquest cas el p-valor de les dues funcions indica que la variable té un efecte important en el model. Cal veure també la diferència d'aquest valor i l'obtingut en la regressió lineal. A la taula 11 es pot veure que l'increment esperat del cost en tenir piscina és de 650.24$

Tornant a una altra variable que fa referència a la superfície de l'habitatge, la variable TOTCSQFT també té valors estimats molt baixos que són semblants entre les dues funcions per ajustar el model. Una vegada més, la desviació típica d'aquesta variable és més baixa en la funció creada que en la funció `rq` tot i que les dues són molt baixes. El p-valors de les dues funcions indíquen que la variable té un efecte important per al model que és que per cada unitat que augmenti aquesta variable, el quantil 0.9 del consum d'energia augmenta 0.103$.

L'última variable que fa referència a una part de superfície de la casa és la variable TOTHSQFT. Per aquesta variable el coeficient estimat és el mateix i indica que per cada unitat que augmenta aquesta variable el quantil 0.9 del cost d'energia augmenta 0.043$. Les desviacions típiques per aquesta variable també són molt baixes tot i que ho és més en la funció `optim`. Respecte als p-valors, la funció `rq` indicaria que es pot considerar aquesta variable no és important per al model mentre que el p-valor indíca que en un nivell de significació del 5% aquesta variable no es considera important, però depenent del mètode d'estimació que es vulgui fer servir sí que podria afectar al model.

A priori el nombre de dormitoris de la casa fa pensar que ha de ser una variable important per al consum d'energia ja que quants més dormitoris, més energia es consumeix. Les estimacions dels coeficients d'aquesta variable per a les dues funcions indíquen que aquesta suposició es certa tot i que són una mica diferents entre elles en el cas del quantil 0.9. Les desviacions típiques, com en la major part de les variables de les que s'ha parlat fins ara, és més alta en la funció `rq` però aquesta diferència en la desviació típica canvia els resultats del contrast d'hipòtesis. En els dos casos, el p-valor és significatiu fet que indica que la variable es consideri important per al model. Per cada habitació que hi hagi, en el quantil 0.9 del consum d'energia augmenta en 71.3$.

L'altra variable que té valors molt diferents entre les funcions per ajusta la regressió quantílica és la variable SOLAR. Per aquesta variable, la funció `rq` dona un valor al coeficient molt més baix que la funció que s'ha creat en aquest treball. Tot i així, la desviació típica segueix sent molt més baixa en la funció `optim`.

L'efecte estimat indica que el fet de produir energia solar fa augmentar el quantil 0.9 del cost d'energia en 90.97$ en el cas de la funció `optim` i en 37.84$ en la funció `rq`. El p-valor per als dos casos és molt elevat i indíca que no hi ha evidències significatives per dir que la variable és important per al model.

Entrant ja a les variables que tenen més relació amb electrodomèstics, és raonable pensar que quant més elevats siguin els valors d'aquestes variables més elevat serà el quantil 0.9 del consum d'energia ja que hi haurà més extrems. En el cas del nombre de televisors de la casa, l'efecte estimat en les dues funcions és pràcticament idèntic i indica que per cada televisor que hi ha a la casa, el quantil 0.9 del consum d'energía augmenta en 141.5$ al quantil 0.9. La desviació típica estimada utilitzant la funció `optim` és més baixa que en l'altra funció però tenint en compte els valors dels coeficients, en ambdos casos la desviació típica és baixa. Els p-valors indíquen que aquesta variable és important per al model.

Un resultat sorprenent per al model és el de la variable que indica el nombre de telèfons mòbils a la casa. És fàcil pensar que quants més mòbils hi han en una casa, la familia és més adinerada i que per tant hi ha en tendéncia un consum més elevat d'energia però els valors estimats per als coeficients indíquen que l'efecte d'aquesta variable és negatiu, és a dir, per cada mòbil que hi ha a la casa, el quantil 0.9 del consum d'energia disminueix en 3.68$. La desviació típica és bastant elevada en comparació amb les estimacions dels coeficients en les dues funcions i això també fa que el p-valor indíqui que la variable no és important per al model. No s'ha trobat gaire sentit a aquest resultat, podria indicar l'absència dels residents i potser una major mobilitat

L'última variable que fa referència a electrodomèstics és el nombre de ventiladors de la casa. Les dues funcions realitzen estimacions similars d'aquest efecte i és més baix del que es pot pensar a priori d'aquesta variable. Per cada ventilador, s'estima que el quantil 0.9 del cost d'energia augmenta en 8.92$. La desviació típica és molt més baixa en la funció `optim` però en les dues funcions aquesta és molt elevada i per això el p-valor indíca que la variable no és important per al model.

Per últim, la variable que indíca el nombre d'habitants a la casa és de les que més efecte té en el consum d'energia. Les dues funcions estimen efectes semblants tot i que la desviació típica segueix sent més baixa en la funció `optim`. El coeficient s'interpreta que per cada habitant que hi hagi a la casa, el quantil 0.9 del consum d'energia augmenta en 151.7$. Les desviacions típiques són molt més baixes que les estimacions dels coeficients en les dues funcions i per això el p-valor indíca que la variable és important per al model.

## 7.2 Assegurances

En aquesta part del treball es presenten els resultats de la regressió quantílica per a la segona base de dades relacionada amb assegurances. Com ja s'ha explicat abans l'elecció del mètode d'optimització i quins valors inicials dels coeficients es prenien per inicialitzar l'optimització, es passa directament a evaluar l'evolució dels coeficients a mesura que augmenta el quantil. A la següent taula es presenten els valors dels coeficients estimats per als diferents quantils.

| Variable | $\beta_{lm}$ | $\beta_{rq0.25}$ | $\beta_{op0.25}$ | $\beta_{rq0.50}$ | $\beta_{op0.50}$ | $\beta_{rq0.75}$ | $\beta_{op0.75}$ | $\beta_{rq0.90}$ | $\beta_{op0.90}$ |
|---|---|---|---|---|---|---|---|---|---|
| Constant | -8120.85 | -2824.98 | -2826.85 | -4588.69 | -4587.25 | -6281.67 | -6287.51 | -6451.74 | -6460.71 |
| lnkm | 1062.54 | 359.81 | 360.16 | 603.55 | 603.40 | 894.77 | 895.21 | 1086.12 | 1086.72 |
| Porc_Vurba | -21.31 | -2.95 | -2.96 | -9.11 | -9.10 | -21.36 | -21.36 | -38.64 | -38.67 |
| Porc_Noctur | 5.35 | 3.18 | 3.17 | 3.49 | 3.52 | 4.07 | 4.13 | 19.69 | 20.05 |
| Edat | 1.37 | -2.77 | -2.81 | -0.52 | -0.53 | 2.42 | 2.46 | 2.19 | 2.32 |
| Sexe | 329.98 | 97.51 | 97.68 | 204.34 | 204.05 | 364.79 | 364.73 | 582.63 | 577.85 |
| Funció Objectiu | - | 2170734 | 2170734 | 4641000 | 4641202 | 3945614 | 3945614 | 2919434 | 2919435 |

**Taula 10:** Efectes de les variables dependents en la regressió quantílica per a diferents quantils en la base de dades d'assegurances.

Com s'observa a la Taula 10, en aquest cas no hi ha pràcticament cap diferència entre els valors dels efectes estimats de la funció `rq` i `optim` i això es veu reflectit en els valors de la funció objectiu que són iguals o en alguns casos la funció `rq` dona valors més baixos. Al estar-se utilitzant exactament els mateixos paràmetres en la funció `optim` que a la base de dades anterior, és possible que els valors de la funció objectiu puguin disminuir més. A continuació es presenten els gràfics de l'evolució dels efectes de les variables.

Observant els gràfics de la Figura 8 es veu que en aquesta base de dades els coeficients que s'han estimat es diferencien molt més dels coeficients del model lineal que els estimats per l'anterior base de dades. Tal i com s'ha vist reflexat a la taula, les dues funcions estimen valors molt similar per a totes les variables en tots els quantils. En el cas de l'efecte constant, en cap moment es pot considerar que sigui igual a l'efecte constant del model lineal. Aquest efecte sempre és negatiu i decreix de forma més o menys constant fins al quantil 0.75. A partir del quantil 0.75 la funció segueix decreixent però ho fa d'una forma menys brusca.

Per a la variable que indica els nombre de kilometres recorreguts l'efecte és creixent durant tots els quantils i bastant constant. Desdel quantil 0.25 fins al 0.75 es pot considerar que aquest és diferent de l'efecte del model lineal. Per al quantil 0.9 l'efecte sí que es considera igual al del model lineal.

Per a la variable que indica el percentatge de kilometres recorreguts durant el dia, l'efecte va decreixent cada vegada més a mesura que el quantil augmenta. Per al quantil 0.25, aquest efecte és pràcticament 0 i a

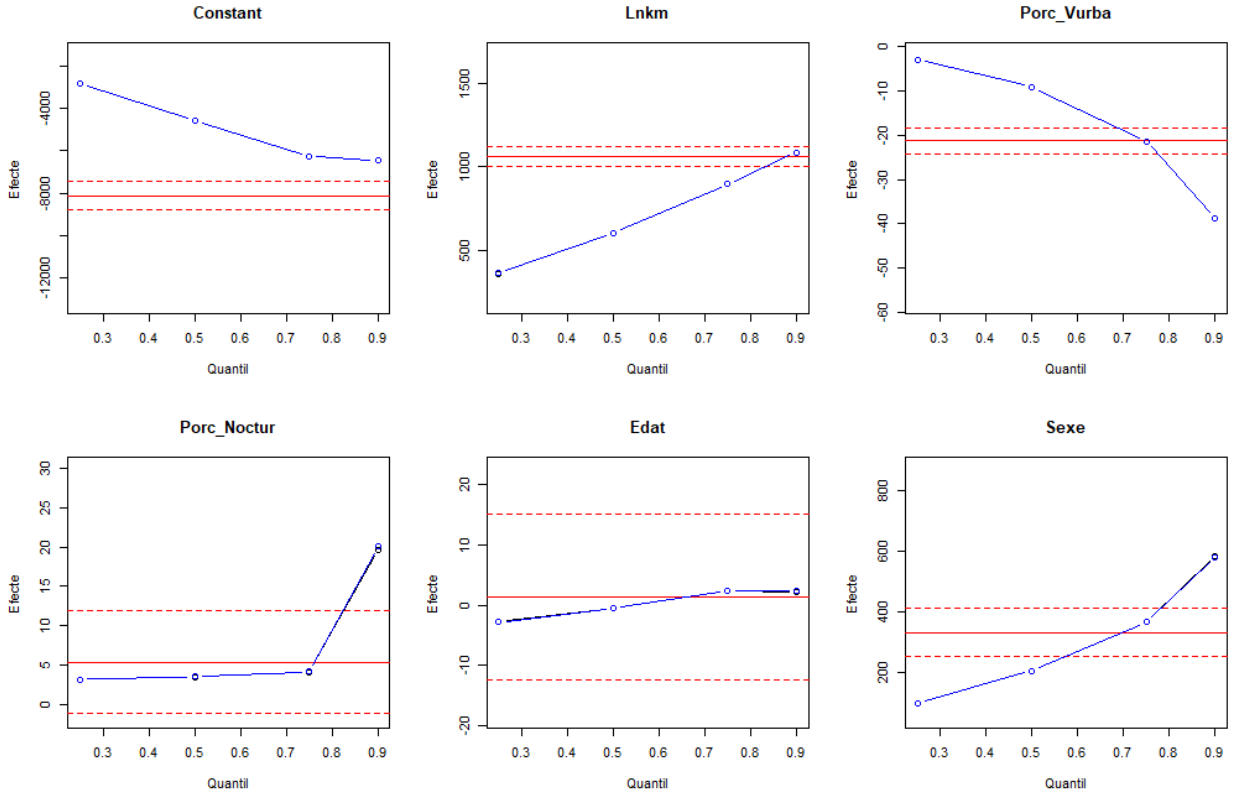


**Figura 8:** Gràfics de l'evolució dels efectes de les variables a mesura que augmenta el quantil en la base de dades d'assegurances.

mesura que augmenta el quantil té un valor negatiu més elevat. Per als quantils 0.25 i 0.5, l'efecte d'aquesta variable fa disminuir menys el quantil de la variable resposta que l'efecte estimat per al model lineal. Per al quantil 0.75, es considera que aquest efecte és igual que el del model lineal i per al quantil 0.9, s'observa que es pot tornar a considerar que l'efecte és diferent al del model lineal.

Per a la variable que indica el percentatge de kilometres conduits de nit l'efecte és practicament el mateix tot i que va creixent lentament a mesura que creix el quantil. A partir del quantil 0.75 aquest creixement augmenta bruscament i s'obté l'únic efecte que es pot considerar diferent que l'efecte del model lineal per aquesta variable.

Per últim, es comenten les dues variables que tenen relació amb les característiques del conductor que són l'Edat i el Sexe. Pel que fa a l'edat, l'efecte pràcticament no varia respecte el quantil 0.25 fins al 0.9 tot i que per als dos primers quantils que s'han estudiat aquesta variable té efecte negatiu i per als quantils 0.75 i 0.9 té efecte postiu. Tot i així en cap dels casos es pot considerar que aquest efecte sigui diferent al del model lineal. En relació a la variable sexe, el comportament és l'invers del de la variable Porc_Vurba, a

mesura que el quantil va augmentant, l'efecte augmenta en major mesura. Per els quantils 0.25 i 0.5 l'efecte és menor que el del model lineal i es pot considerar diferent. Per al quantil 0.75 l'efecte es pot considerar igual i per al quantil 0.9 es pot tornar a considerar diferent i major. De la mateixa manera que en la base de dades anterior, en aquest treball el principal interés es troba en els quantils més elevats i per això s'exploren els resultats per al quantil 0.9 de forma més exhaustiva.

| Variable | $\beta_{rq}$ | $\beta_{optim}$ | $\text{Sd}_{rq}$ | $\text{Sd}_{optim}$ | $\text{p.valor}_{rq}$ | $\text{p.valor}_{optim}$ |
|---|---|---|---|---|---|---|
| Constant | -6451.74 | -6460.71 | 747.77 | 1057.52 | $< 0.001$ | $< 0.001$ |
| Lnkm | 1086.12 | 1086.72 | 58.49 | 89.27 | $< 0.001$ | $< 0.001$ |
| Porc_Vurba | -38.64 | -38.67 | 2.59 | 3.17 | $< 0.001$ | $< 0.001$ |
| Porc_Noctur | 19.69 | 20.05 | 10.01 | 12.76 | 0.049 | 0.117 |
| Edat | 2.19 | 2.32 | 16.70 | 21.57 | 0.900 | 0.910 |
| Sexe | 582.63 | 577.85 | 100.70 | 141.87 | $< 0.001$ | $< 0.001$ |

**Taula 11:** Regressió quantílica al quantil 0.9 de les dades d'assegurances de cotxes

Tal i com s'ha vist anteriorment, els valors dels coeficients estimats són bastant similars entre les funcions però on sí que es poden apreciar canvis importants és en l'estimació de les desviacions típiques. Per a l'efecte constant, s'estima un efecte negatiu molt elevat és a dir, al quantil 0.9 tots els conductors parteixen d'un nombre de kilometres recorreguts per sobre del límit de velocitat negatiu i aquest es va incrementant a mesura que es van afegint els efectes de les variables. Sent més concrets, es parteix de -6460.71 kilometres. La desviació típica estimada a partir de la funció `rq` és molt més baixa que l'estimada a partir de la funció `optim` però en ambdos casos el p-valor és significatiu per tant es tenen evidències estadístiques per considerar que aquest efecte constant és important per al model.

Una altra variable que te un efecte important i de la que és obvi que tindrà un valor elevat del coeficient és el logaritme del nombre de kilometres recorreguts. És lògic pensar que quants més kilometres recorre un conductor, més kilometres recorrerà per sobre del límit de velocitat. La variable explicativa al estar en escala logarítmica s'interpreta diferent que les altres variables. En aquest cas, l'increment d'un 1% en el nombre de kilometres recorreguts fa augmentar el quantil 0.9 en 1086.72 els kilometres recorreguts per sobre del límit de velocitat. Pel que fa a la desviació típica d'aquesta variable la funció `optim` també estima un valor de la desviació típica més alt que en la funció `rq` tot i que és bastant baixa considerant el valor dels coeficients. Aquesta variable també té els dos p-valors significatius per tant es pot considerar que és d'importancia per al model.

Fent referència a per quines zones es solen conduir els vehicles, la variable Porc_vurba estima que el coeficient és de -38.67 i és lògic ja que això indíca que per cada 1% de la distància recorreguda per carreteres urbanes fa disminuir el quantil 0.9 del nombre de kilometres recorreguts per sobre del límit de velocitat. Les

desviacions típiques de les dues variables en aquest cas són bastant similars i els p-valors de les variables donen significatius en ambdues funcions per tant es pot considerar que la variable és important per al model.

Parlant del moment en que els conductors condueixen el seu vehicle, la variable Porc_noctur estima un efecte positiu tot i que és bastant baix. Aquesta variable es pot relacionar amb el fet que els conductors que utilitzen el cotxe per la nit siguin en general gent jove que l'utilitzi per oci i pot portar a pensar que aquests conductors en general tindran valors per sobre del límit de velocitat bastant elevats. La desviació típica d'aquesta variable és alta tenint en compte els valors dels coeficients estimats. Per això, en el cas d'utilitzar la funció `rq` hi ha dubtes a l'hora de considerar aquesta variable com a important per al model ja que el p-valor és molt proper a 0.05 i si es pren com a nivell de significació el 95%, el contrast está just al límit de ser rebutjat. Això no passa en el cas d'utilitzar la funció `optim` ja que el p-valor és més elevat i per tant es consideraria que l'efecte d'aquesta variable no és important en l'ajust. La interpretació del coeficient estimat és que per cada 1% dels kilometres conduits de nit, el quantil 0.9 del nombre de kilometres conduits per sobre del límit de velocitat augmenta en 20.05.

Fent referència a les característiques dels conductors, la variable edat, tal i com s'ha comentat abans, es podria esperar un valor del coeficient negatiu, és a dir, quanta més edat tingui el conductor més responsable serà i això farà disminuir el nombre de kilometres per sobre del límit de velocitat. Es sorprenent que l'efecte estimat per les dues funcions, tot i ser proper a 0, és positiu fet que indicaría el contrari del que es pensava a priori. Per cada any de més que tingui el conductor, el quantil 0.9 del nombre de kilometres per sobre del límit de velocitat augmenta en 2.32. Les desviacions típiques d'aquesta variable són molt elevades en comparació amb els coeficients estimats i per això els p-valors són molt propers a 1 indicant que la variable no es considera important per al model.

Per últim, la variable sexe de la mateixa manera que el nombre de kilometres recorreguts estima un efecte positiu elevat sobre la variable resposta. El valor del coeficient indica que el quantil 0.9 del total de kilometres recorreguts amb excés de velocitat dels homes és 577.85 kilometres superior al de les dones. La desviació típica es baixa tenint en compte el valor tant elevat del coeficient i això fa que el p-valor sigui 0 donant així evidències estadístiques per considerar que la variable sexe és important per al model del quantil 0.9.

# 8 Resultats per al cas inspirat en el TVAR

De la mateixa manera que per a la regressió quantílica s'han realitzat les estimacions dels coeficients per a diferents quantils, també s'han realitzat les estimacions dels coeficients per al cas inspirat en el TVaR a diferents quantils tot i que quan s'utilitza el TVaR es sol fer-ho per a quantils molt elevats. Per a obtenir els resultats s'ha utilitzat el mateix mètode d'optimització amb els mateixos paràmetres i els mateixos valors dels coeficients inicials que per a la regressió quantílica per a un quantil.

## 8.1 Consum d'energia

A la taula següent es presenten les estimacions dels coeficients per als diferents $TVaR_\tau$ juntament amb les desviacions típiques estimades. A part, es mostra el p-valor dels coeficients de les variables per al quantil 0.9 per veure si es consideren útils per a l'ajust o no.

| Variable | $\beta_{lm}$ | $\beta_{0.25}$ | $\text{Sd}_{0.25}$ | $\beta_{0.5}$ | $\text{Sd}_{0.5}$ | $\beta_{0.75}$ | $\text{Sd}_{0.75}$ | $\beta_{0.9}$ | $\text{Sd}_{0.9}$ | $\text{p.valor}_{0.9}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Constant | 512.22 | 516.62 | 132.56 | 547.64 | 95.97 | 700.81 | 110.17 | 222.77 | 435.00 | 0.61 |
| TOTSQFT_EN | 0.031 | 0.021 | 0.016 | 0.01 | 0.024 | 0.038 | 0.027 | 0.17 | 0.05 | 0 |
| SWIMPOOL | 650.24 | 784.35 | 41.89 | 827.16 | 55.39 | 805.35 | 65.05 | 858.50 | 174.40 | 0 |
| TOTCSQFT | 0.12 | 0.14 | 0.018 | 0.12 | 0.017 | 0.12 | 0.017 | 0.073 | 0.03 | 0.015 |
| TOTHSQFT | -0.023 | -0.057 | 0.03 | -0.011 | 0.03 | 0.046 | 0.027 | -0.009 | 0.05 | 0.87 |
| BEDROOMS | 66.14 | 79.78 | 14.75 | 76.0 | 18.89 | 80.43 | 22.31 | 91.84 | 39.50 | 0.02 |
| SOLAR | 179.07 | 124.37 | 119.23 | 207.54 | 90.42 | 176.01 | 105.41 | 928.49 | 432.98 | 0.03 |
| TVCOLOR | 106.04 | 119.36 | 10.23 | 135.84 | 12.98 | 139.82 | 11.57 | 111.17 | 25.40 | 0 |
| NUMSMPHONE | 25.61 | 26.26 | 11.81 | 23.83 | 12.33 | 18.26 | 15.33 | -5.58 | 21.67 | 0.80 |
| NUMFLOORFAN | -1.85 | 8.49 | 11.22 | 6.16 | 10.70 | 5.64 | 15.79 | 10.14 | 23.56 | 0.67 |
| NHSLDMEM | 76.93 | 109.33 | 11.67 | 111.83 | 10.82 | 131.60 | 17.27 | 175.32 | 27.84 | 0 |

**Taula 12:** Regressió quantílica per al cas inspirat en el $TVaR_\tau$ per a les dades de consum d'energía

En la Taula 12 es poden observar els valors estimats dels efectes de la regressió pseudo-quantílica per al cas inspirat en el $TVaR_\tau$ als quantils 0.25, 0.5, 0.75 i 0.9. Seguint la definició del TVaR és normal que a mesura que es va ajustant la regressió per a un quantil més baix, les estimacions s'assemblin més a les del model lineal ja que si el TVaR calcula la mitjana per els valors que són més grans que el quantil $\tau$, si $\tau$ és molt proper a 0 agafarà tota la mostra i per tant es farà regressió sobre la mitjana.

En primer lloc es comentaran els valors dels coeficients per al TVaR al quantil 0.9 i després s'observarà, de la mateixa forma que s'ha fet en la regressió quantílica per al quantil, l'evolució dels coeficients estimats. Per a l'efecte constant s'observa un valor molt més baix que en la regressió quantílica per al quantil 0.9 on es parteix de la base que la gent consumeix 222.77$ en energia prop del $TVaR_{0.9}$. La desviació típica és molt elevada fent que el contrast d'hipòtesis indiqui que el coeficient no és significatiu i que per tant la variable no es pugui considerar important per al model.

Entrant a les variables que descriuen la superfície de la casa, s'observa que per a la variable TOTSQFT_EN el valor del coeficient és molt proper a 0, més concretament 0.17 però en aquest cas la desviació típica d'aquesta variable és molt baixa de tal manera que segons el p-valor, aquesta variable sí que es pot considerar d'importància per al risc d'un elevat consum d'energia a partir del quantil 0.9.

Per a la variable TOTCSQFT, el valor del coeficient també és molt proper a 0 (0.07) però una vegada més la desviació típica també és molt baixa de tal manera que en aquest cas també es pot considerar que aquesta és una variable d'importancia per al model. La interpretació d'aquest coeficient és que per a cada unitat que augmenta la variable TOTCSQFT, a partir del quantil 0.9 el risc d'un elevat consum d'energia s'incrementa en 0.07$.

L'última variable que descriu la superfície d'una part de la casa és la variable TOTHSQFT però en aquest cas, es comporta diferent que les variables de superfície que s'han vist anteriorment. En aquest cas, l'efecte estimat és negatiu fet que implica que per cada unitat que augmentés aquesta variable, el risc d'un elevat consum d'energia disminuiria però aquest valor és practicament 0 i el p-valor per al contrast d'hipòtesis d'aquesta variable indica que no es pot considerar que sigui important per al model. Això també passava quan es realitzava l'estimació per a la regressió quantílica.

Una de les variables que tenia un efecte important per al model per al quantil 0.9 era la variable que indicava si la casa tenia o no piscina. En aquest cas, el valor estimat de l'efecte és molt similar al estimat en el model anterior però la desviació típica és una mica més alta però no és suficient com per que es pugui considerar que es pugui descartar aquesta variable per estimar el model. L'interpretació d'aquesta variable és que el fet de que la casa tingui piscina fa augmentar el risc de tenir una alta despesa d'energia en 858.50$.

La variable que indica el nombre de dormitoris segueix sent important per al model. El valor estimat del coeficient indica que per cada dormitori que tingui la casa el risc de despesa elevada d'energia augmenta 91.84$. Per aquesta variable la desviació típica és molt baixa en comparació amb el valor estimat de l'efecte i és per això que el contrast d'hipòtesi indica que aquesta variable és important per al model.

Un dels canvis més importants que hi ha entre el model per al $TVaR_{0.9}$ i per al quantil 0.9 es troba en la variable SOLAR. En el model per al quantil 0.9 aquesta variable no es considerava important per al model però per al $TVaR_{0.9}$, l'efecte estimat és molt elevat i la desviació típica també però el contrast d'hipòtesis indíca que aquesta variable és important per al model. L'efecte estimat d'aquesta variable és que el fet de produir energia solar incrementa molt el risc de tenir una despesa d'energia elevada en casos extrems (quantil 0.9)

Pel que fa a les variables que descriuen el nombre d'electrodomèstics i aparells electrònics que hi ha a la casa, el nombre de televisors que hi ha a la llar es considera una variable important per al model. Per cada televisor que hi hagi, l'impacte en el risc és de 111.17$. La desviació típica d'aquesta variable és molt baixa comparada amb el valor del coeficient i per això el p-valor és clarament significatiu.

El nombre de telèfons mòbils per altra banda té un efecte negatiu per al model, per cada telèfon el risc de despesa en energia disminueix en 5.58$ però de la mateixa manera que per al quantil 0.9 aquesta variable no es considera important per al model ja que té una desviació típica en l'estimació molt elevada.

Per acabar amb les variables relacionades amb electrodomèstics, el nombre de ventiladors tampoc es considera d'importancia per al model ja que el p-valor tampoc és significatiu. Es torna a tenir una desviació típica molt elevada en comparació amb el valor estimat del coeficient i és el que fa que el p-valor sigui elevat.

Per últim, per al nombre d'habitants de la casa, s'estima que l'augment del risc de despesa elevada d'energia per habitant és de 175.32$. Tot i que és un valor proper al de l'efecte constant, la desviació típica és molt inferior i per això el p-valor dona significatiu per tant hi ha evidències estadísticament significatives per considerar que la variable és d'importància per al model.

A continuació es presenten els gràfics de l'evolució dels efectes de les variables en els diferents nivells. Els valors d'aquests efectes es troben a la Taula 12.

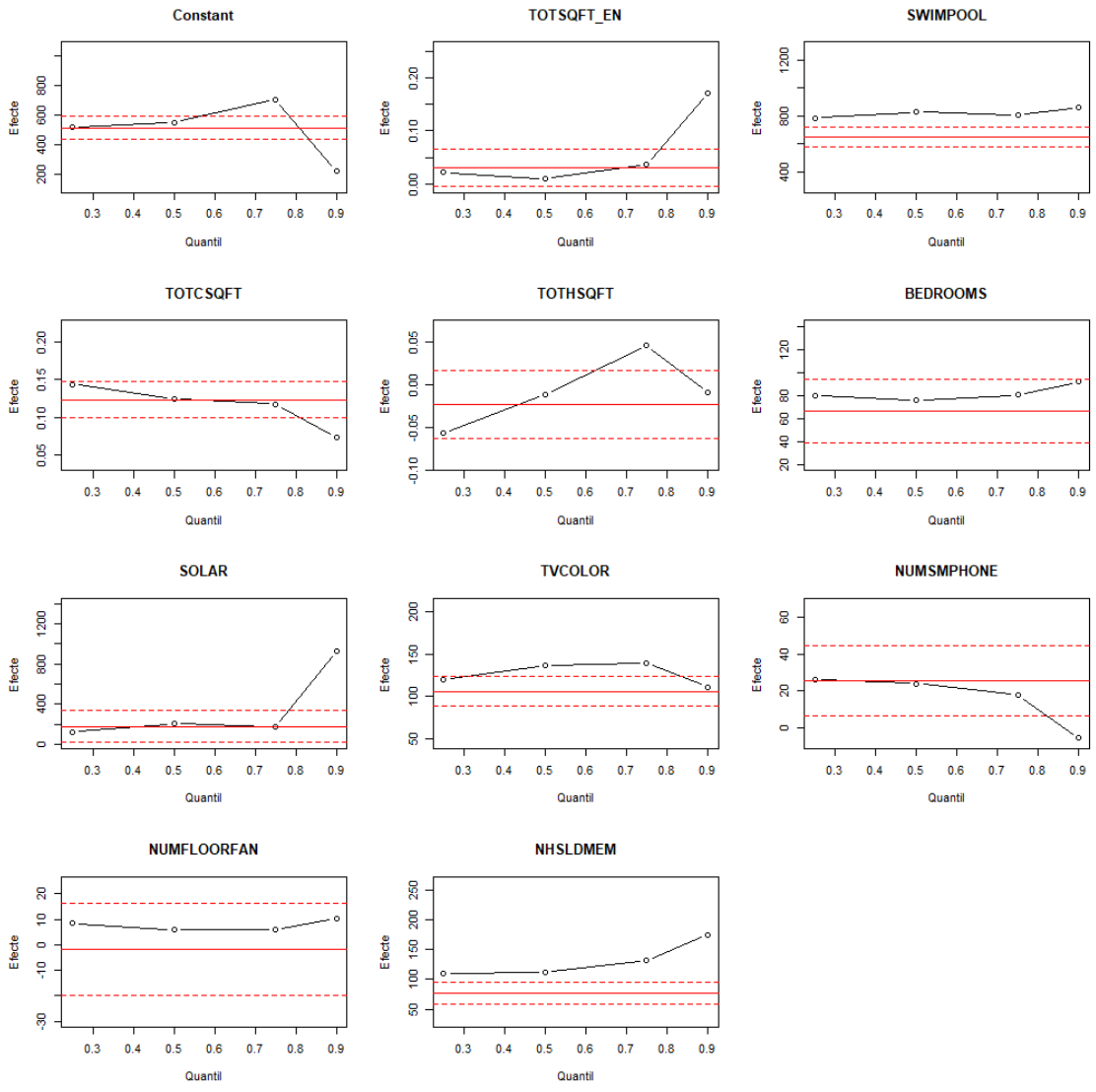


**Figura 9:** Gràfics dels efectes de les variables a mesura que augmenta el quantil a partir del qual es computa el $TVaR_\tau$ en les dades de consum d'energia.

Tal i com s'ha comentat anteriorment és d'esperar que els coeficients del model lineal i els del $TVaR_{0.25}$ siguin semblants i es veu reflexat en què a la majoria de variables l'efecte del $TVaR_{0.25}$ es troba dins l'interval de confiança dels efectes del model lineal i per tant es pot considerar que tenen el mateix efecte.

En el cas de l'efecte constant, s'observa com el TVaR per als quantils 0.25 i 0.5 es troba molt proper a l'efecte del model lineal i per tant es troben dins l'interval de confiança de la regressió lineal. Aquest efecte té un valor més alt en el TVaR per al quantil 0.75 i ja no es pot considerar igual que l'efecte del model lineal. A partir d'aquest punt la tendència canvia i l'efecte per al TVaR al quantil 0.9 és molt més baix que per al quantil 0.75 tant que també es considera diferent que l'efecte del model lineal.

Per a la variable TOTSQFT_EN, s'observa que per als TVaR en els quantils 0.25, 0.5, i 0.75, l'efecte estimat es pot considerar igual que l'efecte del model lineal però no es pot considerar igual per al TVaR en el quantil 0.9. Inicialment l'efecte de la variable es molt proper a 0 i encara s'apropa més des del quantil 0.25 al 0.5. A partir del TVaR per al quantil 0.5 l'efecte va creixent, primer de forma lleugera i després més bruscament, sortint així de l'interval de confiança de l'efecte del model lineal.

L'efecte de tenir o no piscina a la casa és força constant per a tots els TVaR en els 4 quantils. Aquest efecte es troba al voltant dels 800$ i com s'observa en el gràfic es troba sempre per sobre de l'efecte estimat mitjançant la regressió lineal. L'efecte estimat per al TVaR també es troba per sobre de l'interval de confiança de la regressió lineal.

Per a la variable TOTCSQFT, també s'observa com per al TVaR per als 3 primers quantils, l'efecte que s'ha estimat es pot considerar igual a l'efecte estimat mitjançant la regressió lineal. Aquest efecte comença amb un valor al voltant de 0.15 i va decreixent a mesura que s'augmenta el quantil. La diferència de l'efecte és elevada en el TVaR per els quantils 0.75 i 0.9 fent que en el cas del TVaR per al quantil 0.9, es pugui considerar que l'efecte és diferent a l'estimat en el model lineal.

Per a la variable TOTHSQFT, cal recordar que l'efecte estimat per al TVaR al quantil 0.9 era negatiu. L'efecte estimat, tal i com es pot veure en el gràfic, comença sent negatiu en el TVaR per al quantil 0.25 i va creixent a mesura que augmenta el quantil tot i que es manté negatiu per al TVaR al quantil 0.5. En canvi, per al TVaR en el quantil 0.75, l'efecte estimat és positiu i és l'únic en el que l'efecte estimat es troba fora de l'interval de confiança de la regressió lineal. Per al TVaR al quantil 0.9 aquest efecte torna a ser negatiu i tonra a entrar dins l'interval de confiança.

Per al nombre de dormitoris, el comportament de l'efecte de la variable a mesura que augmenta el quantil del TVaR és bastant similar al que s'ha vist en la variable que indicava si hi havia o no piscina. L'efecte és bastant constant tot i que creix una mica del TVaR en el quantil 0.75 al TVaR en el quantil 0.9. Tot i així, l'efecte sempre es troba una mica per sobre de l'efecte del model lineal però mai és suficient com per estar fora de l'interval de confiança del model lineal.

Una de les variables en les que s'ha apreciat un canvi més gran respecte la regressió per al quantil ha estat la variable SOLAR. Aquesta variable tenia un efecte molt elevat per al $TVaR_{0.9}$ mentre que per al quantil, aquesta variable no es considerava important per al model. Com s'observa en el gràfic, per els quantils 0.25, 0.5 i 0.75 l'efecte d'aquesta variable es manté constant i dintre de l'interval de confiança de l'efecte del model lineal però per al quantil 0.9 aquest efecte creix fins a prendre un valor estimat de l'efecte de 928 sent, per descomptat, considerat diferent que l'efecte estimat per al model lineal.

Per al nombre de televisors de la casa, el comportament és bastant peculiar ja que l'efecte estimat per al quantil 0.25 es troba un mica per sota del límit superior de l'interval de confiança del model lineal. L'efecte estimat creix una mica fent que el TVaR per als quantils 0.5 i 0.75 es pugui considerar diferent que l'efecte estimat del model lineal. El que es podria esperar és que l'efecte d'aquesta variable mantingués aquest creixement o fos constant però l'efecte que s'estima per al TVaR al quantil 0.9 és més baix que tota la resta, i s'obté un valor similar al de l'efecte per al model lineal.

Una altra variable que tenia efecte negatiu sobre el cost d'energia en el TVaR era el nombre de telèfons mòbils que hi havia a la casa. Com s'observa en el gràfic, per al TVaR als quantils 0.25, 0.5 i 0.75, l'efecte estimat és positiu i es troba dins l'interval de confiança del model lineal. Es veu que respecte el coeficient estimat per a $TVaR_{0.25}$, aquest efecte és practicament idèntic a l'efecte estimat per al model lineal i va decreixent lleguerament fins a l'efecte estimat per al $TVaR_{0.75}$. De cara al $TVaR_{0.9}$, el decreixement és molt més accentuat i és quan es pren un efecte negatiu que està fora de l'interval del model lineal.

Respecte l'última variable que indica el nombre d'algun tipus d'aparell electrònic es veu que la variable que indica el nombre de ventiladors que hi ha a la casa es comporta de forma similar a l'efecte del nombre de dormitoris. Entre el primer i el segon quantil seleccionats la variable decreix, després es manté constant i per últim acaba creixent però mai se surt de l'interval de confiança del model lineal.

Per últim, el nombre d'habitants de la casa representava una variable important per al model del TVaR en el quantil 0.9 i tal i com es pot observar en el gràfic, sembla que l'efecte també és elevat per a la resta dels quantils seleccionats. L'efecte estimat per el TVaR al quantil 0.25 es troba una mica per sobre dels 100$ per habitant i fora de l'interval de confiança del model lineal. Aquest efecte comença amb un lleuger creixement que es va accentuant a mesura que augmenta el quantil per al qual s'ha ajustat el TVaR.

## 8.2 Assegurances

Un cop estudiats els resultats de l'ajust per al TVaR de la base de dades del consum d'energia, el següent pas és estudiar com s'ajusta la regressió quantílica per al TVaR per a la base de dades d'assegurances de cotxes. Els resultats es presenten de la mateixa manera que per a la base de dades anterior. En primer lloc es mostren els resultats de l'ajust per al TVaR en els quantils $\tau = (0.25, 0.5, 0.75$ i $0.9)$ i s'estudien de forma més exhaustiva els efectes per al quantil 0.9.

| Variable | $\beta_{lm}$ | $\beta_{0.25}$ | $\text{Sd}_{0.25}$ | $\beta_{0.5}$ | $\text{Sd}_{0.5}$ | $\beta_{0.75}$ | $\text{Sd}_{0.75}$ | $\beta_{0.9}$ | $\text{Sd}_{0.9}$ | $\text{p.valor}_{0.9}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Constant | -8120.85 | -5936.59 | 165.93 | -6753.75 | 237.93 | -7050.46 | 394.48 | -6194.13 | 1117.77 | 0 |
| Lnkm | 1062.54 | 831.87 | 16.49 | 953.45 | 20.80 | 1102.14 | 37.44 | 1091.75 | 77.07 | 0 |
| Porc_Vurba | -21.1 | -18.41 | 0.56 | -22.59 | 0.95 | -34.79 | 1.11 | -54.68 | 4.00 | 0 |
| Porc_Noctur | 5.35 | 4.78 | 1.57 | 4.90 | 2.84 | 16.75 | 4.20 | 26.90 | 9.31 | 0.004 |
| Edat | 1.369 | 0.70 | 2.82 | 7.13 | 5.00 | 7.99 | 7.47 | 52.29 | 21.56 | 0.02 |
| Sexe | 329.98 | 285.04 | 19.05 | 369.19 | 30.22 | 526.06 | 47.71 | 907.82 | 116.33 | 0 |

**Taula 13:** Regressió quantílica per al $TVaR_\tau$ per a les dades d'assegurances

En primer lloc per a l'efecte constant s'estima un efecte molt similar al de l'ajust per al quantil 0.9. Aquest efecte és negatiu i es parteix de que per al TVaR al quantil 0.9 els conductors condueixen -6194.13 kilometres per sobre del nivell de velocitat. Tot i que la interpretació sembla no tenir sentit ja que no es pot conduir un nombre negatiu de kilometres, cal recordar que aquest valor només es donaria si la resta de variables fossin 0 i que en aquesta base s'aniran sumant els efectes de les altres variables. La desviació tipica per aquesta variable és molt elevada però no ho és en comparació amb el valor del coeficient. El p-valor del contrast d'hipòtesis indica que aquesta variable es pot considerar important per al model.

Per a la variable que correspon al logaritme del nombre de kilometres que s'ha conduït, tal i com s'ha comentat amb anterioritat és d'esperar que tingui un efecte important per al model ja que quants més kilometres es condueixen, és més fàcil de conduir més estona per sobre del límit de velocitat. Tal i com era d'esperar, el valor de l'efecte estimat és de 1091.75 i la seva interpretació és que per cada 1% que augmenta el valor d'aquesta variable, el risc de tenir un nombre de kilometres conduïts per sobre del límit de velocitat

molt elevat augmenta en 1091.75 en el quantil 0.9. La desviació típica és molt baixa per aquesta variable i el seu p-valor indica que és important per al model.

Una altra variable per a la que es podien fer interpretacions sobre com seria l'efecte estimat era la variable que indica el percentatge de kilometres conduïts per àrees urbanes. Era d'esperar que quant més elevat fos aquest percentatge, més baix sería el nombre de kilometres conduïts per sobre el límit de velocitat ja que per àrees urbanes és difícil de superar aquest límit. Efectivament, l'efecte estimat d'aquesta indica que per cada 1% de conducció per àrees urbanes, el risc en el nombre de kilometres per sobre del límit de velocitat disminueix en 54.68. La desviació típica torna a ser molt baixa i el p-valor torna a indicar que la variable és important per al model.

Per a la variable que descriu el percentatge de kilometres que es recorren de nit s'observa una diferència important entre l'ajust per al quantil i l'ajust per al TVaR. Mentres que en l'ajust per al quantil aquesta variable no era important, en la regressió per al TVaR el coeficient estimat augmenta en gran mesura mentre que la desviació típica disminueix lleugerament. D'aquesta manera, el p-valor per aquesta variable és petit, això porta a concloure que és important per al model. Per cada 1% que es condueixi de nit, el risc en el nombre de kilometres que es recorren per sobre del límit de velocitat augmenta en 26.90.

Entrant en les variables que descriuen els conductors, l'edat també era una variable que no es considerava important per la regressió per al quantil però mentre que la desviació típica es manté constant, l'efecte estimat també augmenta significativament de tal manera que el p-valor del contrast d'hipòtesis passa a ser inferior al 5%. Per cada any que tingui el conductor, el risc en el nombre de kilometres que es condueix per sobre del límit de velocitat augmenta en 52.29. D'aquesta variable podria esperar-se que l'efecte fos negatiu, és a dir quants més anys tingui en conductor, més responsable serà a la carretera però l'efecte estimat indica el contrari. La possible raó és que el conductor tingui una major confiança

Per últim la variable sexe que ja era significativa en el model per al quantil ho segueix sent per a la regressió per el TVaR però l'efecte estimat és molt més elevat que abans. La desviació típica disminueix lleugerament i per això el p-valor segueix sent prou petit. L'interpretació que té aquest coeficient és que el risc dels homes és 907.82 kilometres superior al de les dones en el $TVaR_{0.9}$. Un cop estudiats els efectes de les diferents variables, s'estudia l'evolució dels efectes per els diferents TVaR ajustats.

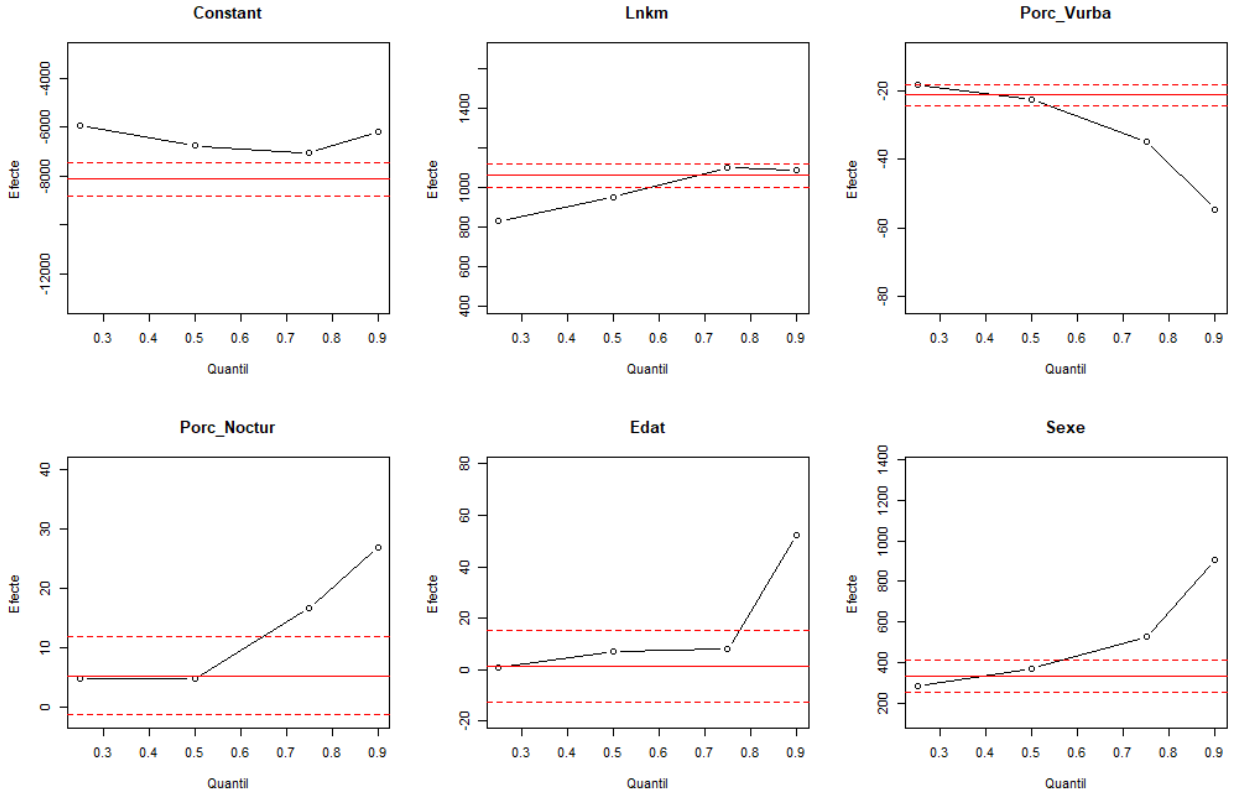


**Figura 10:** Gràfics de l'evolució dels efectes de les variables a mesura que augmenta el quantil a partir del qual es computa el $TVaR_\tau$ en les dades d'assegurances

En els gràfics de la Figura 10 s'observen les diferents evolucions dels efectes de les variables estimats per el TVaR. Per a l'efecte constant s'observa que a mesura que augmenta el quantil del TVaR, l'efecte va sent cada vegada més negatiu i que va apropant-se a l'efecte estimat a partir de la regressió lineal però mai arriba a trobar-se dins de l'interval de confiança del model lineal. En el pas del TVaR per el quantil 0.75 al 0.9 l'efecte de la variable torna a augmentar.

Per a la variable que indica el logaritme del nombre de kilometres conduïts el comportament de l'efecte a mesura que creix el quantil és el contrari de l'esperat. El que s'espera i que ja s'ha explicat abans és que per als primers quantils l'efecte estimat mitjançant la regressió lineal i la regressió quantílica fossin considerats iguals o es trobessin a prop de l'interval i a mesura que anés creixent el quantil aquests efectes es diferenciarien. En el cas d'aquesta variable es pot observar com l'efecte estimat per al TVaR es troba per sota de la regressió quantílica i que va creixent fins al quantil 0.75 on es manté constant i molt proper al límit superior de l'interval. Per aquesta variable l'efecte del model lineal i del quantílic es pot considerar el mateix per als quantils 0.75 i 0.9. Possiblement hi ha un efecte contraposat entre la constant i l'efecte d'aquesta variable, que en aquest cas equivaldria al resultat al model de regressió clàssic.

L'efecte de la variable que indica el percentatge de kilometres conduits per àrees urbanes comença, per als quantils 0.25 i 0.5, dins de l'interval de confiança de l'efecte estimat per al model lineal aquest efecte va sent cada vegada més negatiu i acaba sortint dels límits de l'interval fent que es puguin considerar diferents els efectes dels dos tipus de regressió.

L'efecte que s'estima per la variable que indica el percentatge de kilometres que es condueix de nit té l'efecte contrari. Parteix d'uns valors molt similars als estimats utilitzant la regressió lineal i a partir del quantil 0.5 l'efecte creix a mesura que augmenta el quantil. Per els quantils 0.75 i 0.9 aquesta variable també es pot considerar que és diferent entre els estimats mitjançant els dos tipus de regressió.

La variable edat té un comportament similar al de la variable Porc_Noctur però l'augment de l'efecte tarda més en fer-se. Per el TVaR al quantil 0.25 l'efecte és molt similar al del model lineal i creix una mica per al TVaR al quantil 0.5. L'augment de l'efecte més important es troba en el pas del TVaR per al quantil 0.75 al 0.9 que passa de tenir un efecte de 15 a tenir un efecte de més de 50 de tal manera que es pugui considerar diferent al del model lineal.

Per últim l'efecte de la variable sexe va creixent a mesura que s'augmenta el quantil del TVaR. Per el TVaR al quantil 0.25 aquest efecte es troba lleugerament per sota del l'efecte del model lineal. Aquest efecte va creixent de forma constant fins al quantil 0.75 on l'efecte ja es pot considerar diferent per als dos models. Del quantil 0.75 al 0.9 l'efecte augmenta en gran mesura.

# 9 Conclusions

En aquest treball s'ha estudiat una aplicació de l'eina estadística coneguda com a regressió quantílica. Aquesta eina és poc coneguda i s'inverteix molt poc temps en ensenyar el seu funcionament durant la formació acadèmica ja que es decideix explorar de forma més extensa els mètodes clàssics de regressió com el model lineal. Aquesta eina és útil quan hi ha presència de valors extrems a la variable resposta del model ja que és capaç de tenir-los en compte a l'hora de calcular els efectes de les variables explicatives.

Aquesta aplicació ha consistit en utilitzar la regressió quantílica per ajustar un model inspirat en una mesura del risc coneguda com a TVaR (Tail Value at Risk) també coneguda per altres noms com CTE (Conditional Tail Expectation) o TCE (Tail Conditional Expectation). Aquesta és una mesura del risc que proporciona certs avantatges sobre la mesura del risc més utilitzada que s'anomena VaR (Value at Risk). Mentre que el VaR és centra en comprovar quin és el valor del quantil $\tau$ en una funció d'una variable aleatòria X, el TVaR mira quina és la mitjana de valors que superen el VaR de tal manera que s'obté molta més informació sobre els valors extrems i permet realitzar un millor ajust del risc. Aquesta aplicació va ser tractada en una publicació l'any 2016 i es va demostrar que era possible fer-la però no es va aplicar en casos reals

Abans de començar a ajustar models de regressió, s'ha estudiat quina ha estat l'evolució de la regressió quantílica partint dels primers models de regressió proposats per Boscovich i Laplace que proposaven models on s'intentava minimitzar la suma de desviacions típiques en valor absolut utilitzant sistemes d'equacions. Més endavant Gauss i Legendre parlaven sobre estimar la regressió minimitzant la suma de residus al quadrat que ja era un mètode més similar al que es coneix avuí en dia. Fent lectura de diversos articles que parlen sobre la història de la regressió, es veu que la regressió quantílica és un concepte relativament nou i que s'ha començat a investigar i a aplicar mètodes en les darreres decades deixant així un gran marge d'evolució del mètode.

En aquest treball també s'han explicat diverses eines fora de la regressió quantílica que han estat necessaries per tal d'obtenir resultats. Aquestes són el mètode bootstrap i els contrasts d'hipòtesis. El mètode bootstrap és un mètode de remostreig que s'utilitza sobretot per estimar característiques dels estimadors com serà la variància. Aquest mètode consisteix en agafar un gran nombre de submostres de la base de dades que es vol estudiar i es calcula l'estadístic d'interés per cada una d'elles. A partir d'aquest gran nombre d'estadístics calculats, s'estudia la propietat que sigui d'interés. En el cas d'aquest treball ha estat la desviació típica. Els contrasts d'hipòtesis per altra banda no han tingut un paper tan important com el mètode bootstrap però s'ha considerat oportú fer una molt breu descripció sobre el que són.

Per a la regressió quantílica s'han explicat diversos conceptes que fan que es distingeixi de la regressió lineal. En primer lloc s'ha presentat la funció d'influència. Aquesta és una funció que permet estudiar quin és l'efecte que té una observació sobre una estimació. S'ha plantejat un exemple per tal de comparar l'efecte que tenien les observacions d'una distribució Normal(0,1) sobre l'estimació de la mitjana i de la mediana. Mentre que per a la mitjana l'efecte de la observació era proporcional al seu valor, per a la mediana tots els valors negatius tenien el mateix efecte i tots els positius un altre de manera que es podia deduir que per a valors extrems, l'efecte dels valors anòmals per a la mediana no tenia un gran impacte i que per tant era convenient utilitzar la regressió quantílica.

A continuació també s'ha presentat la funció de pèrdua del quantil que era una funció molt simple que penalitzava el valor de la funció objectiu en funció del quantil sobre el que s'estava fent la regressió i en funció del valor que tenia la condició $(Y - X\beta)$. També s'han mostrat diversos gràfics per tal de comparar els comportaments de les diverses funcions. Com a presentació de les eines utilitzades s'ha acabat mostrant que la regressió quantílica es tracta d'un problema d'optimització i s'ha mostrat quina era la funció objectiu de la regressió quantílica per a un quantil i la funció objectiu de la regressió quantílica inspirada en el TVaR.

Un altre objectiu en aquest treball era utilitzar el software estadístic anomenat R per realitzar aquest treball. Aquest programa ja contenia un paquet de funcions per ajustar la regressió quantílica, en concret la funció `rq` però aquesta no permetia fer-ho per al TVaR que era un dels objectius d'aquest treball. És per això que per tal de cobrir aquesta mancança, s'ha hagut de crear una funció pròpia utilitzant la funció `optim` que serveix per trobar el valor optim d'una funció objectiu. De la funció ja programada s'han comentat els diversos paràmetres dels que es disposava i per la funció creada s'ha comentat el seu funcionament en detall.

La part pràctica d'aquest treball ha estat aplicar la regressió quantílica per els quantils 0.25, 0.5, 0.75 i 0.9 i per als models de riscos basats en el TVaR dels mateixos valors a dues bases de dades diferents. La primera era una base de dades que donava informació sobre famílies dels Estats Units i s'intentava ajustar un model per al consum anual d'energia elèctrica. La segona base de dades provenia d'una asseguradora i donava informació sobre la conducció de vehicles de diverses persones que havien contractat una assegurança de cotxe. L'objectiu d'aquesta segona base de dades ha estat ajustar un model per al nombre de kilometres que els conductors han conduit per sobre del límit de velocitat.

Abans d'aplicar la regressió s'han provat diversos mètodes d'optimització per veure quin d'ells era el que millor valor de la funció objectiu obtenia i s'ha decidit utilitzar el mètode de Nelder-Mead que tot i que tenia un temps de computació bastant elevat, obtenia un bon resultat en el cas de la regressió quantílica. Un altre aspecte important per realitzar l'ajust era decidir quins eren els valors inicials dels coeficients per tal que

la funció `optim` pugués funcionar. Durant la realització d'aquest treball es va veure que depenent de quins valors inicials es donessin, el valor òptim dels coeficients que es trobava era molt diferent. Tenint en compte aquest resultat és molt possible que modificant adeqüadament els paràmetres de la funció `optim` encara s'aconsegueixin millors ajustos per a la regressió però degut a la gran quantitat de temps que comporta aquesta tasca no s'ha realitzat.

També s'ha decidit comparar els resultats obtinguts a través de la funció rq i els obtinguts a través de la funció creada i el resultat ha estat que utilitzant la funció creada es poden aconseguir millors valors de la funció objectiu que la funció rq per a bases de dades amb un nombre bastant gran de variables i observacions però aquesta millora en els valors de la funció objectiu comporta un temps de computació molt més gran que fa que en certs casos sigui més efectiu utilitzar la funció `rq`.

En primer lloc s'ha decidit estudiar els resultats d'aplicar la regressió quantílica per als diversos quantils per tal d'estudiar l'evolució dels efectes a mesura que augmentava el quantil i s'han interpretat amb més detall els del quantil 0.9. Per a la base de dades del consum d'energia s'ha observat que les variables que són importants per a l'ajust són la superfície de la casa, tenir piscina, la superfície del soterrani, parking..., el nombre de dormitoris, el nombre de televisors i el nombre d'habitants de la casa. Totes aquestes variables tenen un efecte positiu és a dir, quant més gran era el valor de la variable més gran és el quantil del consum d'energia hi ha a la casa, per tant el risc de consum extrem augmenta.

Per a la base de dades de les asseguarances de cotxes també s'han estudiat com evolucionaven els efectes de les variables per als diversos quantils i s'han interpretat amb més detall els del quantil 0.9. En aquest cas les variables que eren importants per al model eren el logaritme de la distància total recorreguda, el percentatge de kilometres recorreguts per ciutat i el sexe del conductor. En aquest cas l'efecte d'aquestes variables prenien valors positius.

Un cop estudiats els efectes per als diversos quantils s'ha fet el mateix per ajustar el model inspirat en el TVaR. Al tractar-se d'una mesura de risc té més sentit que s'interpretin els coeficients dels efectes estimats per al quantil 0.9 tot i que també s'ha estudiat l'evolució dels efectes a mesura que s'augmentava el quantil. Per al TVaR hi han hagut canvis en la importància de les variables. La variable que indica si la casa produeix energia solar ha passat a ser considerada d'importancia. Els efectes de totes les variables que són importants per al model sempre són positives. Per tant, incrementen el risc de consum elevat en el cas de les dades del consum d'energia.

Per últim, per la base de dades de les assegurances de cotxes s'ha realitzat el mateix procediment que en les altres parts de resultats i també hi ha hagut canvis en les interpretacions dels efectes de les variables. Per

aquesta segona base de dades, les variables Edat i percentatge de kilometres recorreguts de nit han passat a ser rellevants. Un canvi important ha estat el de l'efecte del percentatge de kilometres recorreguts per ciutat que ha passat a tenir un efecte negatiu és a dir, quant més alt és el percentatge de kilometres recorregut per àrees urbanes menys risc existeix en el total de kilometres que ess recorren per sobre del límit de velocitat.

# 10 Referències


[1] Conde-Amboage, M., Gonzalez-Manteiga, W. & Sanchez-Sellero, C., *Quantile regression: Esimation and lack-of-fit tests*, (2018)

[2] Davino, C., Furno, M. & Vistocco, D., *Quantile regression: theory and applications.*, (2014)

[3] Koenker, R. & Bassett, G. *Regression quantiles*, (1978)

[4] El Bantli, F. & Hallin, M. *$L_1$-estimation in linear models with heterogeneous white noise*, (1999)

[5] Marrocu, E., Paci, R., & Zara, A. *Micro-economic determinants of tourist expenditure: A quantile regression approach. Tourism Management, 50, 13-30.*, (2015)

[6] Liao, W. C., & Wang, X. *Hedonic house prices and spatial quantile regression. Journal of Housing Economics, 21(1), 16-27.*, (2012)

[7] Briollais, L., & Durrieu, G. *Application of quantile regression to recent genetic and-omic studies. Human genetics, 133(8), 951-966.* (2014)

[8] Niemierko, R., Töppel, J., & Tränkler, T. *A D-vine copula quantile regression approach for the prediction of residential heating energy consumption based on historical data. Applied Energy, 233, 691-708.* (2019)

[9] Kaza, N. *Understanding the spectrum of residential energy consumption: a quantile regression approach. Energy policy, 38(11), 6574-6585.* (2010)

[10] Valenzuela, C., Valencia, A., White, S., Jordan, J. A., Cano, S., Keating, J., ... & Potter, L. B. *An analysis of monthly household energy consumption among single-family residences in Texas, 2010. Energy Policy, 69, 263-272.* (2014)

[11] Daniel-Spiegel, E., Weiner, E., Yarom, I., Doveh, E., Friedman, P., Cohen, A., & Shalev, E. *Establishment of fetal biometric charts using quantile regression analysis. Journal of Ultrasound in Medicine, 32(1), 23-33.* (2013)

[12] Behr, A. *Quantile regression for robust bank efficiency score estimation. European Journal of Operational Research, 200(2), 568-581.* (2010)

[13] Tareghian, R., & Rasmussen, P. F. *Statistical downscaling of precipitation using quantile regression. Journal of hydrology, 487, 122-135.* (2013)

[14] Fissler, T., & Ziegel, J. F. *Higher order elicitability and Osband's principle. The Annals of Statistics, 44(4), 1680-1707.* (2016).

[15] Acerbi, C., Szekely, B. *Back-testing expected shortfall. Risk, 27(11), 76-81.*(2014)

[16] Koenker, R. *Quantile regression in R: A vignette. Retrieved November.* (2012)

[17] Koenker, R., Portnoy, S., Ng, P. T., Zeileis, A., Grosjean, P., & Ripley, B. D. *Package 'quantreg'. Cran R-project. org.*(2018)

[18] Hardy, M. R. *An introduction to risk measures for actuarial applications. SOA Syllabus Study Note.* (2006)

[19] https://www.eia.gov/consumption/residential/

[20] Sorenson, H. W. *Least-squares estimation: from Gauss to Kalman. IEEE spectrum, 7(7), 63-68.* (1970)

[21] Kumar, P., & Singh, J. N. *Regression model estimation using least absolute deviations, least squares deviations and minimax absolute deviations criteria. IJCSEE, 3(4), 2320-4028.* (2015)

[22] Rousseeuw, P. J., & Ronchetti, E. *Influence curves of general statistics. Journal of Computational and Applied Mathematics, 7(3), 161-166.*(1981)

[23] Hampel, F. R. *The influence curve and its role in robust estimation. Journal of the american statistical association, 69(346), 383-393.*(1974)

[24] Koenker, R. *Quantile regression: 40 years on. Annual Review of Economics, 9, 155-176.* (2017)

[25] Logan, J. & Petscher, Y. *An introduction to quantile regression* (2013)

[26] Wei, Y., Pere, A., Koenker, R., & He, X. *Quantile regression methods for reference growth charts. Statistics in medicine, 25(8), 1369-1382.* (2006)

[27] Koenker, R., Hallock, K. *Quantile regression: An introduction. Journal of Economic Perspectives, 15(4), 43-56.* (2001)

[28] Droge, B. *Phillip Good: Permutation, parametric, and bootstrap tests of hypotheses. Metrika, 64(2), 249-250* (2006)